PATENT
Docket No.: **19226/2051 (R-5655)**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Thomas A. Szyperski	Examiner: Y. Gakh
Serial No.	:	09/897,583	
Cnfrm. No.	:	1224	Art Unit: 1743
Filed	:	June 29, 2001	
For	:	METHOD OF USING REDUCED DIMENSIONALITY NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY FOR RAPID CHEMICAL SHIFT ASSIGNMENT AND SECONDARY STRUCTURE DETERMINATION OF PROTEINS	

DECLARATION OF THOMAS A. SZYPERSKI UNDER 37 C.F.R. § 1.132

Mail Stop: RCE
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

I, Thomas A. Szyperski, pursuant to 37 C.F.R. § 1.132, declare:

1. I received a Diploma degree in Chemistry from Technical University of Munich, Germany in 1988 and a Dr. Sc. degree in Chemistry from ETH Zurich, Switzerland in 1992.

2. I am currently Professor of Chemistry and Biochemistry, and Director of the UB High-Field NMR Facility at University at Buffalo, The State University of New York, Buffalo, New York. I am also currently Adjunct Senior Researcher at the Hauptman-Woodward Medical Research Institute, Buffalo, New York.

3. As indicated in my attached Curriculum Vitae (Exhibit 1) and list of publications (Exhibit 2), I have authored or co-authored over 90 peer-reviewed professional publications in the fields of nuclear magnetic resonance (NMR) techniques and structure

determination of biological macromolecules using NMR spectroscopy. Since 1999, I have given over 90 invited lectures in these same technical fields (see Exhibit 1).

4. I am an elected member of the American Chemical Society, American Association for the Advancement of Science, and the Gesellschaft Deutscher Chemiker.

5. I am the inventor of the above-identified patent application.

6. I am presenting this declaration to demonstrate that, contrary to the statement on page 4 of the outstanding office action for the above-identified patent application, none of the NMR experiments disclosed in Fernández et al., "NMR With ^{13}C, ^{15}N-Doubly-Labeled DNA: The *Antennapedia* Homeodomain Complex With a 14-mer DNA Duplex," *J. Biomol. NMR* 12:25-37 (1998) ("Fernández"), Gehring et al., "H(C)CH-COSY and (H)CCH-COSY Experiments for ^{13}C-Labeled Proteins in H_2O Solution," *J. Magn. Reson.* 135(1):185-193 (1998) ("Gehring"), Yamazaki et al., "Two-Dimensional NMR Experiments for Correlating $^{13}C\beta$ and ^{1}H δ/ε Chemical Shifts of Aromatic Residues in ^{13}C-Labeled Proteins Via Scalar Couplings," *J. Am. Chem. Soc.*, 115:11054-11055 (1993) ("Yamazaki"), Schirra, "Three Dimensional NMR Spectroscopy" http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/3dnmr.htm (1996) ("Schirra"), or "Cell Cycle/Gene Regulation," http://daisy.bio.nagoya-u.ac.jp/golab/pdb/pdb2nmb.txt (1998) ("Cell Cycle Protocol") discloses the same conditions created by the reduced dimensionality (RD) NMR experiments disclosed in Szyperski et al., "Sequential Resonance Assignment of Medium-Sized $^{15}N/^{13}C$ -Labeled Proteins with Projected 4D Triple Resonance NMR Experiments," *J. Biomol. NMR* 11:387-405 (1998) ("Szyperski"), that is, a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum. In addition, I am presenting this declaration to show that the present invention satisfied a long-felt, but unfulfilled need. Finally, I am presenting this declaration to show that, at the time the present invention was made, there was considerable disbelief and skepticism in the field of NMR spectroscopy with regard to whether my method would be an effective method to use for rapid protein structure determination.

None of the Experiments Disclosed in Fernández, Gehring, Yamazaki, Schirra, or Cell Cycle Protocol Creates the Same Conditions Created by the Reduced Dimensionality NMR Experiments Disclosed in Szyperski

7. I am familiar with the disclosures of Szyperski, Fernández, Gehring, Yamazaki, Schirra, and Cell Cycle Protocol.

8. Szyperski describes the use of projected four-dimensional (4D) triple resonance NMR experiments for the efficient sequential resonance assignment of $^{15}N/^{13}C$-labeled proteins, where a reduced dimensionality (RD) three-dimensional (3D) $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN NMR experiment is recorded either in conjunction with a RD 3D HNN<CO,CA> NMR experiment or with a RD 3D HNNCAHA NMR experiment. As described on pages 388-392 and in Figures 1-2, the RD NMR experiments disclosed in Szyperski involve a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in an n-1 dimensional spectrum, thereby encoding one of the n chemical shifts into the spectral separation of the two peaks.

9. Fernández teaches obtaining ^{1}H, ^{13}C, and ^{15}N NMR assignments for a doubly-labeled 14-base pair DNA duplex in solution, both in the free state and complexed with the uniformly ^{15}N-labeled *Antennapedia* homeodomain. The resonance assignments are obtained in three steps: (i) identification of the deoxyribose spin systems via scalar couplings using 2D and 3D HCCH-COSY and soft-relayed HCCH-COSY; (ii) sequential assignment of the nucleotides via ^{1}H-^{1}H nuclear Overhauser effects (NOEs) observed in 3D ^{13}C-resolved NOESY; and (iii) assignment of the immino and amino groups via ^{1}H-^{1}H NOEs and ^{15}N-^{1}H correlation spectroscopy. Fernández describes conventional multidimensional NMR experiments, where the indirect chemical shift evolution times sample only a single chemical shift, as indicated by the radiofrequency pulse schemes (shown in Figure 2) used in these experiments. The conditions created in RD NMR, i.e., a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum, can be achieved only if a second chemical shift is jointly sampled with the one being detected in quadrature. No such joint sampling is even implemented or suggested in Fernández. Thus, contrary to the statement in the outstanding office action, the conditions created by the conventional NMR experiments disclosed in Fernández are <u>not</u> the same conditions created by the RD NMR experiments disclosed in Szyperski. Accordingly, there can be no basis for modifying the experiments disclosed in

Fernández to conduct a RD NMR experiment as taught by Szyperski. Moreover, it is not trivial to identify the best choice for two chemical shifts to be jointly sampled in order to conduct a RD NMR experiment, and Fernandez provides no information in that respect. Fernández neither attempts to nor achieves the assignment of protein resonance by conducting any RD NMR experiments, let alone a RD 3D H,C,C,H-COSY NMR experiment. Furthermore, Fernández does not provide any expectation that conducting a RD NMR experiment would be useful in obtaining protein resonance assignments.

10. Gehring discloses three NMR experiments for identifying carbon and proton sidechain resonances in ^{13}C-labeled proteins. The first experiment is an improved H(C)CH-COSY experiment comprising the application of gradients for coherence selection and a reduction in the phase cycle. The second experiment is a new (H)CCH-COSY experiment with two carbon dimensions. The third experiment is a 2D proton-edited (H)C(C)H-COSY experiment that allows suppression of methylene resonances. Gehring describes conventional multidimensional NMR experiments, where the indirect chemical shift evolution times sample only a single chemical shift, as indicated by the radiofrequency pulse schemes (shown in Figure 1) used in these experiments. The conditions created in RD NMR, i.e., a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum, can be achieved only if a second chemical shift is jointly sampled with the one being detected in quadrature. No such joint sampling is even implemented or suggested in Gehring. Thus, contrary to the statement in the outstanding office action, the conditions created by the conventional NMR experiments disclosed in Gehring are not the same conditions created by the RD NMR experiments disclosed in Szyperski. Accordingly, there can be no basis for modifying the experiments disclosed in Gehring to conduct a RD NMR experiment as taught by Szyperski. Moreover, it is not trivial to identify the best choice for two chemical shifts to be jointly sampled in order to conduct a RD NMR experiment, and Gehring provides no information in that respect. Gehring neither attempts to nor achieves the assignment of protein resonance by conducting any RD NMR experiments, let alone a RD 3D H,C,C,H-COSY NMR experiment. Furthermore, Gehring does not provide any expectation that conducting a RD NMR experiment would be useful in obtaining protein resonance assignments.

11. Yamazaki discloses two-dimensional NMR experiments, (Hβ)Cβ(CγCδ)Hδ and (Hβ)Cβ(CγCδCε)Hε, for correlating ^{13}Cβ and ^{1}H δ/ε chemical shifts of aromatic residues in ^{13}C-labeled proteins based on scalar connectivities. Yamazaki describes conventional two-dimensional NMR experiments, where the indirect chemical shift evolution times sample only a single chemical shift, as indicated by the radiofrequency pulse schemes (shown in Figure 1) used in these experiments. The conditions created in RD NMR, i.e., a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum, can be achieved only if a second chemical shift is jointly sampled with the one being detected in quadrature. No such joint sampling is even implemented or suggested in Yamazaki. Thus, contrary to the statement in the outstanding office action, the conditions created by the conventional NMR experiments disclosed in Yamazaki are not the same conditions created by the RD NMR experiments disclosed in Szyperski. Accordingly, there can be no basis for modifying the experiments disclosed in Yamazaki to conduct a RD NMR experiment as taught by Szyperski. Moreover, it is not trivial to identify the best choice for two chemical shifts to be jointly sampled in order to conduct a RD NMR experiment, and Yamazaki provides no information in that respect. Yamazaki neither attempts to nor achieves the assignment of protein resonance by conducting any RD NMR experiments, let alone a RD RD 2D HB,CB,(CG,CD),HD NMR experiment. Furthermore, Yamazaki does not provide any expectation that conducting a RD NMR experiment would be useful in obtaining protein resonance assignments.

12. Schirra discloses several 3D NMR experiments and their magnetization transfer pathways. Schirra describes conventional multidimensional NMR experiments, where the indirect chemical shift evolution times sample only a single chemical shift. The conditions created in RD NMR, i.e., a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum, can be achieved only if a second chemical shift is jointly sampled with the one being detected in quadrature. No such joint sampling is even implemented or suggested anywhere in Schirra. Thus, contrary to the statement in the outstanding office action, the conditions created by the conventional NMR experiments disclosed in Schirra are not the same conditions created by the RD NMR experiments disclosed in Szyperski. Accordingly, there can be no basis for modifying the experiments disclosed in

Schirra to conduct a RD NMR experiment as taught by Szyperski. Moreover, it is not trivial to identify the best choice for two chemical shifts to be jointly sampled in order to conduct a RD NMR experiment, and Schirra provides absolutely no information in that respect. Schirra neither attempts to nor achieves the assignment of protein resonance by conducting any RD NMR experiments, let alone any RD 3D NMR experiment. Furthermore, Schirra does not provide any expectation that conducting a RD NMR experiment would be useful in obtaining protein resonance assignments.

13. Cell Cycle Protocol is a copy of a printout of a protein databank entry, listing a number of conventional multidimensional NMR experiments that were recorded. Cell Cycle Protocol describes conventional multidimensional NMR experiments, where the indirect chemical shift evolution times sample only a single chemical shift. The conditions created in RD NMR, i.e., a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in a n-1 dimensional spectrum, can be achieved only if a second chemical shift is jointly sampled with the one being detected in quadrature. No such joint sampling is even implemented or suggested anywhere in Cell Cycle Protocol. Thus, the conditions created by the conventional NMR experiments disclosed in Cell Cycle Protocol are <u>not</u> the same conditions created by the RD NMR experiments disclosed in Szyperski. Accordingly, there can be no basis for modifying the experiments disclosed in Schirra to conduct a RD NMR experiment as taught by Szyperski. Moreover, it is not trivial to identify the best choice for two chemical shifts to be jointly sampled in order to conduct a RD NMR experiment, and Cell Cycle Protocol provides absolutely no information in that respect. Cell Cycle Protocol neither attempts to nor achieves the assignment of protein resonance by conducting any RD NMR experiments, let alone any RD 3D or 2D NMR experiment. Furthermore, Cell Cycle Protocol does not provide any expectation that conducting a RD NMR experiment would be useful in obtaining protein resonance assignments.

The Present Invention Satisfied a Long-Felt, But Unfulfilled Need

14. The method of the present invention which uses a suite of RD NMR experiments on a protein sample to obtain rapid and complete protein resonance assignment for protein structure determination satisfied a long-felt, but unfulfilled need.

15. Rapid resonance assignment is a prerequisite for rapid protein NMR structure determination and, thus, for high-throughput (HTP) structure determination and structural genomics. The aims of structural genomics have been to (i) explore the naturally occurring "protein fold space" and (ii) contribute to the characterization of function through the assignment of atomic-resolution three-dimensional (3D) structures to proteins. The ultimate goal is to provide one or more representative 3D structures for every structural domain family in nature. It is now generally acknowledged that NMR will play an important role in this endeavor. The resulting demand for HTP structure determination requires fast and automated NMR data collection and analysis protocols.

16. Two key objectives for NMR data collection can be identified. Firstly, the measurement time should be minimized so as to lower the cost per structure and relax the constraint that NMR samples need to be stable over long time periods. Secondly, automated analysis requires recording of a redundant set of NMR spectra each affording good resolution, while it is also desirable to keep the total number of spectra small to reduce complications due to interspectral variations of chemical shifts. This second objective can be addressed by maximizing the dimensionality of the spectra. However, the joint realization of the first and second objective is impeded by the large lower bounds for measurement times of four (or higher) dimensional NMR spectra arising from the independent sampling of three (or more) indirect dimensions.

17. "Sampling limited" and "sensitivity limited" data collection regimes can be distinguished, depending on whether the sampling of the indirect dimensions or the sensitivity of the multidimensional NMR experiments *per se* determines the minimally achievable measurement time. Because structure determinations rely on nearly complete shift assignments routinely obtained using ^{13}C,^{15}N,^{1}H-triple-resonance (TR) NMR, the development of techniques that avoid the sampling limited regime has been an important challenge.

18. The fact that there had been a long-felt need for rapid and complete protein resonance assignment was recognized by those of ordinary skill in the art, as shown by a number of publications published in 2001. For example, Heinemann et al., "High-Throughput Three-Dimensional Protein Structure Determination," *Current Opinion in Biotechnology* 12:348-354 (2001) (attached hereto as Exhibit 3) stated the following:

> The NMR structure determination process in itself consists of a number of different time consuming steps, independent of sample preparation, which until recently could take up to several months or even years. Recording a data set took a minimum of six to eight weeks, and the assignment of resonance signals, including sidechain signals, required at least three to four weeks. Furthermore, the interpretation of the NOEs (nuclear Overhauser effects) observed in NOESY (NOE spectroscopy)-type spectra could take a couple of months, followed by one or two weeks of structure calculation. To achieve high-throughput, these times have to be significantly reduced (emphasis added).

See id. at 351.

In another 2001 journal article, Prestegard et al., "Nuclear Magnetic Resonance in the Era of Structural Genomics," *Biochemistry* 40:8677-8685 (2001) (attached hereto as Exhibit 4), Prestegard et al. noted the following:

> A more severe limitation was that the time required for NMR data acquisition and analysis is long, and sample preparation requires the use of isotopically labeled media (^{15}N- and ^{13}C-labeled proteins). There have been enormous strides made in the efficient production of proteins through expression in *E. coli* (citation omitted), and new cell-free production techniques pioneered in Japan promise more latitude in produced proteins and incorporated labels (citation omitted). However, the 4-6 weeks of acquisition and subsequent months-long periods required for assignment and structure determination is still a major obstacle (citation omitted). This time scale is not compatible with structural genomics objectives that would require 100-200 structures per year from each of the seven NIH-sponsored pilot centers (citation omitted) (emphasis added).

See id. at 8680. Thus, at the time the present invention was made, about 4 to 8 weeks of NMR instrument time per protein structure were considered to be a realistic estimate for ~1 mM protein samples with molecular weights up to 15 kDa and, as indicated by the above statements, there had clearly been a long-felt and unfulfilled need to achieve "rapid" and complete protein resonance assignment.

19. The failure to solve this long-felt need was, in part, due to the lack of appreciation of the present invention's potential to accomplish rapid and complete protein resonance assignment. Thus, despite the advantages of RD NMR in reducing sampling requirements and minimal measurement time and my prior publications describing a number of specific RD NMR experiments, there had been a general lack of interest in the field in using RD NMR for rapid protein resonance assignment. This is well demonstrated by the attached Panel Summary for a grant proposal entitled "Reduced Dimensionality NMR Spectroscopy for Structural Genomics" that I submitted to the National Science Foundation (NSF) when I sought to have my initial work on the present invention funded. The Panel Summary, which was attached to a letter sent by Acting Deputy Division Director Christopher Greer, Ph.D. on May 5, 2000 (attached hereto as Exhibit 5), indicating that the grant proposal could not be supported, outlines the salient points raised in the panel discussion of my proposal. It stated:

> The PI proposes to further develop methods to reduce the time for structure determination by NMR. This work is based on some nice experiments that the PI has already published. The PI projects that implementation of an RD NMR package would result in time savings for a facility trying to maximize throughput for protein NMR. Unfortunately, the project is essentially for the generation of software with no development of science. <u>The NMR community has not appeared so far to be interested in RD.</u> The panel concluded that new demonstrations of practicality of RD NMR were needed (emphasis added).

20. The subject matter of the present invention was published by, *inter alia*, me in Szyperski et al., "Reduced-Dimensionality NMR Spectroscopy For High-Throughput Protein Resonance Assignment," *Proc. Natl. Acad. Sci. USA* 99:8009-8014 (2002) (a copy attached hereto as Exhibit 6). Comments from two reviewers for the journal were attached to the

acceptance letter from the PNAS Office (attached hereto as Exhibit 7). In describing the aspects of the paper that are novel and important, Reviewer #II, stated:

> This paper describes the generalization of the concept of "reduced dimensionality" and its application to the NMR pulse sequences for protein resonance assignment. Efficiency in NMR data collection is one of the important concerns especially in structural genomics projects. This is not only a proposal of a concrete answer but also a proposal suggestive to other general problems (emphasis added).

In addition, Reviewer #I noted the following:

> The reduced dimensionality experiments which are discussed in this paper are very important for enhancing information content to enable automated assignment of NMR resonances, and thereby high throughput structure determination by NMR. The concept of reduced dimensionality experiments is not new, but this manuscript puts together (for the first time to the knowledge of this reviewer) a suite of experiments. which are specifically useful for automated data analysis. The particular set of experiments seems to have been selected to be compatible with the AUTOASSIGN program from one of the authors, other sets may be optimal in other contexts but what is presented demonstrates the concept very well (emphasis added).

21. From these statements from referees in a high-quality journal such as the Proceedings of the National Academy of Sciences, it is apparent that the method of the present invention constitutes a significant advance in the art by having the capability to provide rapid and complete resonance assignment for high-throughput protein structure determination.

Disbelief and Skepticism of Those Skilled in the Art Regarding the Present Invention's Effectiveness

22. At the time the present invention was made, there was considerable disbelief and skepticism in the field of NMR spectroscopy with regard to whether my method would be an effective method to use for rapid protein structure determination.

23. Structure determinations of proteins by NMR rely on the nearly complete assignment of chemical shifts, which can be obtained using multidimensional ^{13}C,^{15}N,^{1}H-triple-resonance NMR methods. At the time I demonstrated rapid and complete backbone and side chain resonance assignments for the "Z domain" of *Staphylococcal* protein A by conducting a suite of RD ^{13}C,^{15}N,^{1}H-triple-resonance NMR experiments, it was unknown whether conducting a suite of RD NMR experiments on a protein sample would be a sufficiently efficient method for rapid protein structure determination. Instead, there was prominent and prevalent concern that the added spectral overlap in the RD NMR spectra would render the use of RD NMR experiments ineffective in obtaining protein resonance assignments for structure determination. Thus, at the time the present invention was made, although the general concept of RD NMR had been previously published and would offer an attractive solution to reduce the minimal measurement times, no one had demonstrated obtaining "rapid" and complete resonance assignments for a protein by conducting a suite of RD NMR experiments. To the contrary, considerable skepticism existed in the field with regard to whether RD NMR would be an effective method to use for protein structure determination. Moreover, as already noted above in paragraph 19, there had been a general lack of interest in the field in using RD NMR for rapid protein resonance assignment.

24. There were a number of reasons why no one believed that my method would be an effective method to use for protein structure determination. Firstly, NMR spectroscopy was relatively insensitive, which severely limited experimental design. Typically samples at ~1 mM protein concentration were required, preventing studies of proteins with very low solubility. Because of constraints on pulse sequence design arising from these sensitivity limitations, several different NMR spectra recorded over a four to six week period were necessary to obtain the information needed for a high-quality structure determination. These long data collection periods, in turn, put significant constraints on sample stability. Although multiple samples can be used in the structure determination process, each one must be stable for days to weeks with respect to precipitation, aggregation, and other forms of degradation. Secondly, manual analysis of these multiple NMR data sets was laborious and required significant expertise. Finally, in analyzing the NMR data, the density of constraints was sometimes inadequate for accurate structural analysis.

25. The considerable disbelief in the field regarding the present invention is demonstrated by the attached reviews of a grant proposal that I submitted to the NSF when I sought to have my initial work on the invention funded and by the attached reviews of a manuscript that I submitted to the Journal of American Chemical Society when I sought to have my work on the invention published. As the reviewers' comments reveal, both the reviewers of the NSF and the Journal of American Chemical Society shared strong skepticism that rapid and complete protein resonance assignment could be obtained by conducting a suite of RD NMR experiments and that my method would work on larger size proteins. Nevertheless, my NSF grant proposal was supported shortly after an initial rejection, while my manuscript, after being rejected by the Journal of American Chemical Society for containing subject matter too specialized for the general readers of the journal, was eventually published in a similar prestigious journal which has an even more generalized readership than the Journal of American Chemical Society. Furthermore, the initial view by experts in the field that my method would not be effective in determining the structure for larger size proteins was subsequently refuted by a number of publications co-authored by me.

26. After submitting a grant proposal entitled "CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics" to NSF, I received a letter dated November 9, 1999, from Acting Deputy Division Director Jerry D. Cohen, Ph.D. (attached hereto as Exhibit 8), initially indicating that the grant proposal could not be supported. Attached to that letter were reviews of the grant proposal by different reviewers and a Panel Summary. The Panel Summary stated:

> While recognizing that the PI pioneered the RD method in NMR, the panel believed that the PI must address important issues raised by the reviewers. In particular, the panel questioned whether the RD method will be broadly applicable, particularly to larger proteins with congested spectra. In addition, the PI should address the extent to which RD approaches truly reduce the main bottlenecks in NMR structure determination for high throughput structural genomics. The PI should address in more detail possible pitfalls in the proposed approaches and suggest alternatives.

In particular, one of the reviewers stated:

> (1) It is not clear that this approach will benefit the structural genomics initiative. In proteins where spectral overlap is an issue, the introduction of more crosspeaks in the RD-NMR spectrum will pose a real limitation to the usefulness of this approach.

A second reviewer, in discussing the subject matter of the proposal, noted the following:

> This CAREER proposal involves the development and application of reduced-dimensionality NMR to protein structure determination with the long-range goal of contributing to structural genomics. A focus will be on reducing the amount of NMR time required for resonance assignment and structure determination of proteins. Dr. Szyperski is one of the original developers of reduced dimensionality methodology when he was a postdoctoral in Wuethrich's [*sic*] lab and thus is uniquely qualified to carry out the proposed research program....The proposal is well written and the experiments are well outlined. An added strength of the proposal is the collaboration with Dr. Guy Montelione, where Dr. Montelione's AUTOASSIGN program will be use modified to incorporate the reduced dimensionality data. This is a potentially important simplification of the bottleneck in protein structure determination, the resonance assignments.

A third reviewer noted the following:

> The PI has been the leader in the development of "reduced dimensionality" NMR experiments. In this approach, n chemical shifts are encoded into nû1 [*sic*] dimensions of a multidimensional NMR spectrum in order to maximize digital resolution while minimizing acquisition time.
>
> * * *
>
> I am most enthused about the proposed search for a robust minimum set of NMR experiments for structure determination of small protein domains and proteins. Defining such sets will be important for high throughput structure determination in structural genomics initiatives. The collaboration with Prof. Montelione also will enable the selected experiments to be integrated with the strengths of the AUTOASSIGN program. The proposal would be stronger if more discussion was presented of the criteria to be used in recognizing an appropriate set of

> experiments. How will the PI show that the designed set of experiments is generalizable (to proteins outside the test set) and robust (resistant to common experimental difficulties (poor resolution, disorder, etc.)?
>
> * * *
>
> The potential widespread application of the proposed methods in structural genomics provides the broad impact envisioned by the NSF's review criterion 2.

A fourth reviewer stated:

> This application by Dr. Thomas A. Szyperski, proposes to develop approaches to more efficiently and more rapidly elucidate protein structures by using NMR techniques. The proposal addresses both fundamental research and applied research targeting the commercial efficiency of NMR groups focused on structural genomics initiatives.
>
> * * *
>
> Dr. Szyperski proposes to use a set of three proteins (ubiquitin, 8.6 kDa; RNaseA, 15 kDa; and Ns-1, 17 kDA). These three proteins are not representative of most proteins that will come from the genome. What about proteins in the 20 kDa to 30 kDa (or higher) where TROSY experiments promise to have the greatest impact in NMR, especially for its use in structural genomics? This has not been addressed. What are the limitations of his eventual protocol for using NMR in structural genomics programs.

A fifth reviewer stated:

> 1) At the highest level, Dr. Szyperski needs to address if NMR is the proper method to participate in the process of structural genomics. This premise is just stated in the third line of the project summary and the third paragraph of the background section, but it is not supported by discussion at all. Can NMR really compete with X-ray for rapid turn over of structures? If so, on what subset of structures? What spectral properties are required?
> 2) Depending on the outcome of this question, the next issue should be a discussion of where the bottleneck for NMR structure determination lies.

This reviewer is convinced that it does not lie with the resonance assignment at all. The most time consuming step by far is the NOE identification and iterative structure calculation and refinement. This holds for small proteins as well as large ones. There has even always been a upper size limit for which assignments could still be obtained, but no structure. Therefore, improvement of the efficiency of spectral assignment step does not really reduce the time necessary for structure determination and is therefore of very limited use to structural genomics.

3) At a third level, this reviewer differs strongly with Dr. Szyperski's assessment on the efficiency of reduced dimensionality experiments. The experiments are claimed to be more efficient than the current suite of existing experiments. This can be arguably be the case for very small proteins for which no overlap exists in the spectra. However, for the smallest of proteins there is no assignment problem at all, and new methods are not necessary either.

The reduced dimensionality approach places more peaks in the NMR spectrum, which is always a bad idea when spectra get complicated. The next problem is that in experiments such as HNCAHA, there are two Ca frequencies per HN. Thus, the reduced dimensionality experiment will generate 4 peaks for these two peaks for which it is not known how they pair up. It gets worse if there is 2D HN degeneracy. Dr. Szyperski's solution (published by him in the past) to overcome this is to mistune the tau_4 delay in order to obtain axial peaks that contain the HNCA information. But, mistuning will negatively affect the sensitivity of the HNCAHA peaks. The next problem is that in order to obtain high resolution in N-1 dimensions, the indirect time- domain FIDs needs to be collected to high resolution. This leads to low sensitivity of the data. Also, the RD N-1 dimensional experiments cause a splitting of all peaks, which cause a reduction of sensitivity as compared to the N-dimensional experiment. At best, the sensitivity becomes equal if spectra get symmetrized around the center CA positions (which is different for every amino acid). As such, this reviewer is not really happy with the further development of the reduced dimensionality experiments.

27. Subsequent to the submission of the above career grant proposal, I submitted a significantly revised grant proposal entitled "Reduced Dimensionality NMR Spectroscopy for Structural Genomics" to NSF. On May 5, 2000, Acting Deputy Division Director Christopher Greer, Ph.D. sent me a letter (see Exhibit 5), indicating that the grant proposal could not be supported. Attached to that letter was a Panel Summary, as well as

reviews of the proposal by different reviewers (attached hereto as Exhibit 9). In describing the significance of the proposed work, one of the reviewers stated:

> Criterion 1: Dr. Szyperski proposes further develop reduced dimensionality (RD) triple resonance NMR methods for use in structural genomics. The research will focus upon development of a protocol that will allow identification of minimal sets of NMR data for structure determination. These advances will be incorporated into automated structure determination schemes to further enhance applicability to a large number of protein targets. The major strengths of the proposal are the expertise of the P.I. in the general area and the importance of the time reductions which potentially could result from the work thereby facilitating determination of a large number of protein structures.
>
> * * *
>
> Criterion 2: The development of methods to facilitate the rapid solution of atomic resolution structures will have a profound impact on the area of structural genomics. Clearly, one of the limitations of structure determination with high-field NMR is the time required for data collection. The studies proposed hold the potential to significantly lessen this time constraint and to lead to the applicability of modern NMR methods to a larger number of proteins.

A second reviewer noted the following:

> This is a very good - excellent proposal by a starting investigator who has made excellent contributions in the field of biomolecular NMR. The proposal is very sound and promises to greatly enhance the utility of solution NMR for structural genomics. For many of the proteins that are likely to be considered by NMR reduced dimensionality NMR is the way to go, offering savings in measuring time. Thus the development of a robust set of experiments is an important step. The applicant has made very important contributions in this area previously and he has the equipment, resources and expertise to continue. I recommend funding.

A third reviewer stated:

The PI has proposed an orchestrated approach for addressing one of the three major "bottle necks" in the use of high resolution NMR in structure-based genomics. He proposes the development of innovative and novel techniques that would reduce the total instrument time required for obtaining the data needed for sequential resonance assignments and structure determination for small to medium sized proteins. This work dovetails nicely into major research initiatives in the area of functional genomics that are being pursued both here in the US and elsewhere. The PI is superbly qualified and equipped for the proposed studies, with a strong publication record in this area and with ample computer facilities and NMR instrumentation at his disposal. The proposed studies would make important and relatively novel contributions to the field of structural genomics by developing techniques in two general areas. First, further development of reduced dimensionality (RD) experiments have the potential to dramatically reduce the time required for acquisition and improve the quality of multidimensional NMR experiments for rapid assignment and structure determination. Second, the development of techniques to combine the measurement of residual dipolar couplings with resonance assignment experiments, and to use the variations in residual dipolar couplings to resolve chemical shift degeneracy will further optimize the amount of information extracted from experimental data.
Potential problems seem to be adequately addressed. The most obvious problem in the RD approach is of course the loss of spectral resolution. As the PI points out, the RD triple resonance techniques will benefit from incorporation of TROSY schemes, since the slowly relaxing component selected by TROSY yields a sharper resonance peak and will be of great value in optimizing spectral resolution. This should be adequate for the small to medium sized proteins to which the proposed techniques would be applied.

A fourth reviewer noted the following:

This is a very good application of a talented scientist who recently joined the faculty of SUNY Buffalo. The proposed work is to utilize and further develop a principle called "reduced dimensionality" (RD) for reducing the measurement time of multidimensional NMR experiments. The PI developed the method while he was in the laboratory of Dr. Wuthrich in Zurich. The idea behind this approach is to record data in a way that both the sum and the difference of the frequencies of a pair of spins is measured along a single indirect dimension so that the information of two dimensions can be read in a

single dimension. Thus, one can obtain the information of a 4D experiment essentially in the time one usually spends for a 3D experiment. In more recent implementation, Dr Szyperski has developed a way to also record the central peak (axial peak), which may provide additional useful information. The PI claims that this is particularly useful for applications in structural genomics where it is important to acquire spectral information as fast as possible. The proposal also proposes to use this methodology for measuring residual dipolar couplings in partially aligned systems. The PI describes convincingly that this information can be obtained from the central peaks in the more recent RD pulse sequences.
This is a technique-oriented proposal. The PI has first described the RD principle seven years ago. Several other groups have applied the technique although it hasn't yet found wide-spread use. However, the benefit of shorter measuring times claimed by the PI is obvious. The reason why the technique hasn't had wide impact may be that it requires additions and subtractions of peak positions to obtain chemical shift data, which complicates analysis of large data sets. This will be less of an issue if the experiments are incorporated in automated assignment routines. Thus, the proposed collaboration with the Montelione laboratory is a very positive aspect of this proposal. Another reason why the RD approach hasn't been used widely is that the primary limitation of NMR structure determination is still the process of making well-behaving protein samples. This may no more be a concern in a structural genomics effort where hopefully many well-behaving proteins will await there structures being solved. In this respect, the choice of ubiquitin and protein Z and other well-characterized proteins is a little distracting.
Overall, this is a very good application by a new investigator who is a highly talented NMR expert. The technology development proposed has potentially high impact for protein structure determination. The knowledge and research of the PI will have high educational impact on the local structural biology community.

A fifth reviewer noted:

This proposal focuses on expanding the role of high resolution NMR in a very important, and currently high profile, area of scientific activity, structural genomics. Hypothesis underlying this general area is that the wealth of information coming from sequencing projects can be tapped by solving sufficient numbers of protein structures quickly to provide examples from all fold families (several thousand). X-ray crystallography is clearly the major player in this area, but NMR is

> important because of its applicability to proteins that do not give diffraction quality crystals. One primary limitation of NMR is that data collection using conventional approaches is slow, requiring on the order of a month of spectrometer time for each protein to be solved. This proposal would expedite the process by identifying a minimal set of NMR experiments, basing these on reduced dimensionality experiments, exploring suitability for automated assignment, and extending schemes to measurement of residual dipolar couplings.
> The activity proposed is very useful and it would be carried out under the direction of an investigator with an excellent record. The reduced dimensionality trick, which relies on a proper collection of zero and two quantum coherences, is novel, and one pioneered by the investigator (although there are other examples using simultaneous evolution of two types of chemical shift to reduce dimensions). The proposed interaction with the Montelione group on incorporation of the reduced dimensionality experiments into the AUTOASSIGN program is excellent. And, the extension to collection of residual dipolar data could do much to improve the reliability of assignment and structure determination.

28. Despite the favorable and positive comments by the different reviewers as shown in the preceding paragraph, the grant proposal was rejected, as indicated by the May 5, 2000, letter from Christopher Greer, Ph.D mentioned in the preceding paragraph; however, shortly after, Thomas E. Smith, Ph.D. from the same Division mailed me a letter (attached hereto as Exhibit 10), informing me that my proposal would be supported.

29. Subsequently, I submitted a manuscript entitled "Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis" to the Journal of American Chemical Society to have my work on the invention published. On July 17, 2001, Associate Editor Dr. F. Ann Walker mailed me a letter (attached hereto as Exhibit 11), rejecting the manuscript for publication and attaching comments from two reviewers. In that letter, Dr. Walker stated that both of the reviewers feel that the work was appropriate for publication in the Journal, but only after a number of major points are addressed. In particular, Reviewer 1 stated:

> This paper is largely a statement of advocacy rather than a critical scientific evaluation and account. While there is certainly merit in

reduced dimensionality experiments for accelerating resonance assignments, the case study of a 4.5. ns correlation time polypeptide, the projected enhancements with cryoprobes, single transient spectra, etc. are not only misleading for the general scientific readership but scientifically indefensible with the data shown.

* * *

Thus, the paper would only be acceptable after major revision. This includes shortening the paper to remove much of the redundancy in statements and in clearly delineating the results from predictions.

Reviewer 2, in discussing the subject matter of the manuscript, noted the following:

The manuscript of Szyperski et al. presents a suite of reduced dimensionality triple resonance experiments for the rapid assignment of the backbone and side chain resonances of small to medium-sized proteins. A major part of the discussion deals with a critical comparison of the sensitivity of the individual experiments and the optimisation of measuring time taking into account the requirements of high spectral resolution and sufficient signal to noise. In addition, analysis of these NMR data has been implemented in the program package AutoAssign. Although I believe that the manuscript contains some interesting ideas, I am sceptical [*sic*] whether it merits (in the present form) publication in the Journal of the American Chemical Society .

The authors present eight unpublished pulse sequences for triple resonance experiments using the so called reduced dimensionality approach. These experiments are well presented with a comprehensive description of the different pulse sequences in the Supporting Information. This part of the manuscript is certainly of interest to the biomolecular NMR community and merits publication in a specialised NMR journal.

The major aim of the manuscript, however, is to present a new general strategy for speeding up the NMR assignment step of proteins. Reduced dimensionality experiments provide connectivities between four different nuclei and the high spectral resolution required for an automated data analysis. In addition, the authors present a comparison of the sensitivity of eleven NMR experiments recorded on a small protein of 63 residues at a magnetic field strength of 14.1 T, and discuss theses results in terms of "minimal required data set" and "minimal required experimental time". Although I believe that NMR assignments will greatly benefit from the use of reduced

dimensionality experiments, I have some major concerns about the conclusions drawn from the experimental results. The conclusions about the relative sensitivity of the NMR experiments are certainly valid for other proteins with similar molecular weight (well below 100 residues) and studied under similar experimental conditions (temperature, magnetic field strength), but they will (as the authors agree on page 13) completely change for larger molecules (higher tumbling correlation times). Along the same lines, the chosen set of eleven triple resonance experiments may (or may not) be the best choice for assignment of a small protein, but other experiments will certainly yield much better results in the case of larger perdeuterated or randomly fractional deuterated proteins. Finally, a combination of reduced dimensionality experiments (for the most sensitive ones) and standard 3D triple resonance experiments (for the less sensitive ones) seems to me a better choice than the one presented in the manuscript. Thus the major conclusions of the manuscript are rather subjective and far from being general. By the way, the statement that complete NMR assignment of medium sized molecules will be possible "within a day or less" (page 19) is a very optimistic statement. It could even have a negative impact on most of the biomolecular NMR laboratories, where this step still requires a couple of weeks up to several months. This sentence should therefore be dropped from the manuscript unless the authors prove that they are really capable of what they are claiming.

In conclusion, for the manuscript to be acceptable for publication in JACS, I suggest that the authors add at least one additional experimental study on a second test molecule (in the range of about 150 residues). This will help the work to be of more general interest to the readership of JACS. As recording of the NMR data sets and assignment can be accomplished in a couple of days by the authors, this requirement should not significantly delay publication of this work.

30. On August 8, 2001, I mailed a revised version of the manuscript to the Associate Editor of the Journal of American Chemical Society accompanied by a letter (attached hereto as Exhibit 12) which responded to the reviewers' comments in detail. In responding to the reviewers' criticisms, I revised the manuscript to avoid claims beyond what was actually shown in the paper.

31. On August 13, 2001, the Associate Editor of the Journal of American Chemical Society sent me a letter (attached hereto as Exhibit 13), indicating that her journal remained unable to publish my manuscript due to the comments by Reviewer 1, to which my

revised manuscript was sent. Reviewer 1 was of the view that the revised manuscript was not appropriate for publication in the journal, but rather should appear in a more specialized journal.

32. I soon after responded to the Associate Editor of the Journal of American Chemical Society (attached hereto as Exhibit 14), addressing each of the criticisms made by Reviewer 1 and pointing out the inappropriateness of all of the reviewer's points. In that letter, I asked whether it would be possible to receive the opinion of the Reviewer 2 on the revised manuscript.

33. On September 6, 2001, the Associate Editor of the Journal of American Chemical Society mailed me a letter (copy attached hereto as Exhibit 15), informing me that my manuscript was sent back to Reviewer 2, as well as a new reviewer who was an independent expert in the field, Reviewer 4, and, based on the reviewers' evaluations of the manuscript, concluding that the manuscript was not appropriate for publication in her journal. In particular, Reviewer 2 made the following comment:

> The revised manuscript of Szyperski et al. has significantly improved with respect to the original version by deleting much of the idealized projections to future achievements. This manuscript thus merits publication in a journal with a readership largely interested in the practical details of biomolecular NMR. The question remains whether it merits publication in the Journal of the American Chemical Society. Triple resonance experiments as well as reduced dimensionality spectroscopy are well known concepts widely used for the resonance assignment of proteins and nucleic acids. The detailed description of eleven RD triple-resonance experiments is certainly not of much interest to the broad readership of JACS. What would be of interest to many readers of JACS, especially molecular biologists interested in NMR as a tool to resolve structures, study molecular interfaces, etc., is the experimental proof that a certain set of NMR experiments combined with an automated assignment protocol will yield rapid assignment for a wide range of proteins. To achieve this goal, the proposed concept has to be applied to different (at least two or three) molecular systems in different molecular weight ranges and eventually using different isotope labels (e.g. partial deuteration). It is not sufficient to apply the method to a single small protein and add a reference to a couple of other proteins studied using the same strategy without providing any experimental evidence. By the way the spectrum of Fig. S20 also looks to me to correspond to a rather

unstructured protein. Thus at the current state the proposed strategy is just one of many different possible strategies. I am actually using a different one based on another set of experiments, and neither me nor T. Szyperski and coworkers have proofed so far that their strategy is the more efficient one. The scientific content does not gain much by putting the work in the context of structural genomics. A valuable strategy to speed up the assignment process will be of interest to any NMR spectroscopist working on proteins not only those involved in structural genomics. On the other side the work would certainly benefit from a more thorough analysis of other time limiting factors like data processing, peak picking, and automated assignment.
In agreement with Reviewer 1 I believe that the manuscript lacks sufficient experimental proof that the conclusions are valid for a large range of molecular systems, although some of the more specific points raised by Reviewer 1 can be argued to be rather subjective and of minor importance.
In conclusion, I suggest that the major part of the manuscript containing the detailed analysis of the RD experiments and their application to the Z-domain should be published without significant changes in a more specialised NMR journal such as J. Biomol. NMR, and that the authors eventually resubmit at a later date a manuscript which discusses in more experimental detail their assignment strategy with respect to other methods for the example of representative protein systems without the need of a detailed description of NMR pulse sequences.

Reviewer #4, noted the following:

After reading the paper and associated correspondence, I agree with the Editor's decision. Both reviewers were initially of the opinion that the paper was not suitable for J.A.C.S., and the second review by Reviewer 1 reiterated this position. It seems to me that the points raised in the final correspondence from the author (disputing the Editor's decision) are not really the most important factors in the decision, but constitute more of a philosophical difference between the author and the reviewer. The main issue is that the subject of the paper is too specialized for the general readership of J.A.C.S. Publication in J.Biomol. NMR or even J.Magn.Reson. would be a much more appropriate option for this paper, since the target audience is actually a very small subset (those interested in rapid and automated NMR resonance assignments of proteins) of a small subset (those interested in NMR assignments of proteins) of a subset (those interested in biological NMR) of readers of J.A.C.S.

34. After the final rejection of my manuscript by the Journal of American Chemical Society, I submitted a substantially similar manuscript to the Proceedings of the National Academy of Sciences, which is as prestigious and has an even more generalized readership than JACS. On April 19, 2002, I received a letter from the PNAS Office (see Exhibit 7), informing me that the PNAS Editorial Board accepted my manuscript for publication. My published paper was cited in a recent article in the Journal of American Chemical Society by one of the leading research groups in the field (attached hereto as Exhibit 16), where the RD approach was hailed as being one of the most successful methods that have been proposed for rapid data collection. Subsequent work by me and my co-workers, where the same or substantially similar suite of RD NMR experiments was applied on different proteins of sizes ranging from 8 kDa to 21 kDa, was recognized and validated in additional papers published in Journal of Biomolecular NMR (attached hereto as Exhibits 17, 18, 19, and 20), Proteins (attached hereto as Exhibits 21, 22, and 23), and Protein Science (attached hereto as Exhibits 24 and 25). For example, Szymczyna et al., "Letter to the Editor: ^{1}H, ^{13}C, and ^{15}N Resonance Assignments and Secondary Structure of the PWI Domain From SRm160 Using Reduced Dimensionality NMR, *J. Biomol. NMR* 22:299-300 (2002) (see Exhibit 17) shows that the total acquisition time for obtaining assignments of the protein backbone and side chain atoms for a 12.5 kDa polypeptide containing the PWI motif of SRm160 using a suite of RD NMR experiments was 44.5 hours.

35. The above-described series of events demonstrates that, at the time the present invention was made, there was substantial skepticism regarding whether it would be effective in obtaining rapid and complete resonance assignment for protein structure determination and whether it could be successfully applied to larger size proteins. Nevertheless, as demonstrated by the acceptance of my paper in the Proceedings of the National Academy of Sciences and subsequent publications by me and my co-workers, I was able to demonstrate to the people in the field that my invention was effective in achieving this result.

36. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date: 11/21/2005

Thomas Szyperski

Thomas A. Szyperski

EXHIBIT 1

Curriculum Vitae

Dr. sc. nat ETH Thomas Szyperski
Full Professor
Director, High-field NMR facility
The State University of New York at Buffalo
Chemistry Department
('cross appointed' in Departments of Biochemistry and Structural Biology)
816 Natural Sciences Complex
Buffalo, NY 14260, USA

Date of birth	December 13, 1963
Place of birth	Berlin, Germany
Nationality	German
Civil status	Married, three children
Obligatory Military Service	July 1 1982 - September 30, 1983

Education

Vordiplom Biochemistry, University of Tübingen, Germany	1984
Diplom Chemistry ('1.2'), Technical University München, Germany	1985
Vordiplom Chemistry ('1.0'), Technical University München, Germany	1988
Dr. sc. nat. ETH, ETH Zürich, Switzerland	1992
Mentor: Prof. K. Wüthrich	
Habilitation, ETH Zürich, Switzerland	1998

Employment History

TU München, Germany	3.1988 - 9.1988
University of Auckland, New Zealand	10.1988 - 2.1989
ETH Zürich, Switzerland	3.1989 - 10.1998
State University of New York, USA	11.1998 –

Other Positions

Member of

Operating Committee of the New York Structural Biology Center
(http://www.nysbc.org/)

Leader of

NMR Division of the Northeast Structural Genomics Consortium
(http://www.nesg.org/

Awards

7.1981 / 7.1982	Member of the german national team participating at the XIII. and XIV. International Chemistry Olympiads - awarded Silver and Golden Medal, respectively.
5.1993	'Medaille der ETH' for outstanding dissertations
7.1998	'Privatdozent' ETH Zürich
11.1999	Research Innovation Award of the Research Corporation
3.2003	Buck-Whitney Medal of the American Chemical Society
11.2003	*Scientific American* 50 Award for 'Chemistry and Material Sciences'

Honors

9.1982 - 9.1988	Scholar of the 'Studienstiftung des Deutschen Volkes'
6.1989 - 9.1991	Scholar of the 'Verband der Chemischen Industrie'
5.1999 -	Adjoined Senior Researcher at 'Hauptman-Woodward Medical Institute'

Professional Memberships and Activities

Member of

American Chemical Society
Gesellschaft Deutscher Chemiker
American Association for the Advancement of Science

Member of

Faculty of 1000

Professional Service

Member of

Editorial board of the 'Journal of Structural and Functional Genomics'

Reviewer for scientific journals:

Applied and Environmental Microbiology
Biochemistry
Biopolymers
Bioinformatics Journal
Biotechnology and Bioengineering
Biotechnology Progress
BioTechniques
European Journal of Biochemistry
Journal of the American Chemical Society
Journal of Bacteriology
Journal of Biomolecular NMR
Journal of Magnetic Resonance
Journal of Molecular Biology
Journal of Structural and Functional Genomics
Macromolecules
Magnetic Resonance in Chemistry
Metabolic Engineering
Nature Biotechnology
Structure

Community Service

Reviewer for

NIH: BBCA study section (2000)
NIH: S10 Instrumentation Panel (2004)
NSF: Major Research Instrumentation Panel (2001)
NSF: *Ad Hoc* Reviewer (2002 - 2005)
NSF: Panel Member (2004, 2005)
Wellcome Trust, United Kingdom (2001; 2003)
Genome Canada (2004)
Wiener Wirtschafts-, Forschungs- und Technologiefonds (2005)
Israel Science Foundation (2005)

Entrepreneurial Activities

Co-founder of

'Metabolic Concepts GmbH', Zürich (founded 1998 as an ETHZ spin-off consulting company)

Member of

Scientific Advisory Board of 'GeneFormatics' (2001-2002)

International Patents

Method of using G-matrix Fourier transformation nuclear magnetic resonance (GFT NMR) spectroscopy for rapid chemical shift assignment and secondary structure determination of proteins. US patent number 6.831.459.

Two patents on methodology for rapid sampling of NMR data are pending.

Invited Lectures (1999 -)

1.USA

(1) "NMR Spectroscopy: a Powerful Tool for Life Scientists"
Biochemistry Department, State University of New York at Buffalo
Buffalo, New York, February 15, 1999.

(2) "Sequential Resonance Assignment of Medium-sized $^{15}N/^{13}C$-Labeled Proteins with Projected 4D Triple Resonance Experiments"
Varian NMR Users Conference
Orlando, Florida, March 1, 1999.

(3) "Indirect Detection of ^{13}C using 1D and 2D [^{13}C, ^{1}H]-correlation NMR Spectroscopy.
NIH Symposium: ^{13}C in Metabolic Research, University of Texas Medical Branch
Dallas, Texas, May 6, 1999.

(4) "NMR Spectroscopy in Structural Biology"
Center of Advanced Research in Molecular Biology and Immunology
Buffalo, New York, May 18, 1999.

(5) "NMR Spectroscopy Applied in Structural Biology and Metabolic Research"
Hauptman-Woodward Medical Research Institute
Buffalo, New York, June 10, 1999.

(6) "NMR at SUNYAB"
1st Upstate New York NMR Symposium, Wadsworth Center
Albany, New York, October 4, 1999.

(7) "Reduced Dimensionality NMR Experiments for Structural Genomics",
Symposium NE Structural Genomics Consortium Project Team, Rutgers University,
Piscataway, New Jersey, November 2, 1999.

(8) "Structural Biology of the Mitochondrial Origin of Light Strand DNA Replication" Rockefeller University, New York, New York, December 13, 1999.

(9) "Reduced Dimensionality NMR Spectroscopy for Structural Genomics"
Northeastern Structural Genomics Consortium, Rutgers University,
Piscataway, New Jersey, December 14, 1999.

(10) "Novel RD NMR Experiments for Structural Genomics"
NESG Consortium Workshop,
Princeton, New Jersey, May 13, 2000.

(11) "METAFoR by NMR: an Approach Comes of Age"
Cargill Dow Polymers,
Minnetonka, Minneapolis, May 24, 2000.

(12) "Reduced-dimensionality NMR for Structural Genomics"
Pacific Northwest National Laboratories,
Redland, Washington, July 29, 2000.

(13) "Structural Biology in Supercooled Water"
2nd Upstate New York NMR Symposium, Cornell University,
Ithaca, New York, October 2, 2000.

(14) "News on RD NMR"
Cornell Medical School
New York, New York, November 28, 2000.

(15) "Reduced-dimensionality NMR spectroscopy"
RD NMR workshop UB high-field NMR facility and CCR Buffalo,
Buffalo, New York, December 5, 2000.

(16) "Structural Genomics by NMR"
Foster Chemistry Colloquia UB Chemistry Department
Buffalo, New York, December 8, 2000.

(17) "Reduced-dimensionality NMR Spectroscopy: An Approach Comes of Age"
Keystone Symposia, Frontiers of NMR in Molecular Biology
Big Sky, Montana, January 22, 2001.

(18) "Bio-NMR at UB: Supercool!"
Chemistry Department, Youngstown State University
Youngstown, Ohio, February 9, 2001.

(19) "Structural Biology in Supercooled Water"
UB Physics Department Seminar Series
Buffalo, New York, February 13, 2001.

(20) "Structural Biology in Supercooled Water"
Varian NMR Users Conference
Orlando, Florida, March 9, 2001.

(21) "Structural Genomics by NMR"
Center for Computational Research 2000-2001 Colloquium Series
Buffalo, New York, March 27, 2001.

(22) "Metabolic Flux Ratio and Bioreaction Network Topology Analysis by NMR"
Department of Plant Biology Michigan State University
East Lansing, Michigan, April 25, 2001.

(23) "Metabolic Flux Ratio and Bioreaction Network Topology Analysis by NMR"
Microbia Inc.
Boston, Massachusetts, April 27, 2001.

(24) "Structural Biology in Supercooled Water"
3rd Upstate New York NMR meeting
Rochester, New York, October 15, 2001.

(25) "Metabolic Profiling by NMR"
Industrial Associates Program New Jersey
Princeton, New Jersey, October 16, 2001.

(26) "RD NMR for Structural Genomics"
Bristol-Myers-Squibb
Princeton, New Jersey, October 17, 2001.

(27) "Structural Biology in Supercooled Water"
NJ American Chemical Society NMR Topic Group
Princeton, New Jersey, October 17, 2001.

(28) "Metabolic Profiling by NMR" Metabolic Profiling: Pathways in Discovery Cambridge Healthtech Institute's Premier Conference
Chapel Hill, North Carolina, December 4, 2001.

(29) "Structural Genomics by NMR"
NIEHS
Chapel Hill, North Carolina, December 5, 2001.

(30) "NMR at UB"
Kent State University
Kent, Ohio, January 31, 2002.

(31) "NMR Sample Preparation for High-throughput Structure Determination"
NIH-PSI workshop
Bethesda, Maryland, March 7, 2002.

(32) "NMR-based structural genomics"
Buffalo-Niagara Post-genomic Research Conference
Buffalo, New York, May 15, 2002.

(33) "New NMR methods for Structural Genomics"
NESG consortium, Annual Meeting Center for Advanced Biotechnology and Medicine
Piscataway, New Jersey, June 10, 2002.

(34) "News on RD NMR"
NIH workshop University of Wisconsin
Madison, Wisconsin, June 21, 2002.

(35) "Metabolic Profiling: New Insights"
2002 Annual Meeting of the Society for Industrial Microbiology
Philadelphia, Pennsylvania, August 11-15, 2002.

(36) "NMR Methodology for High-Throughput Protein Resonance Assignment"
Biotechnology Forum
Buffalo, New York, October 22, 2002.

(37) "NMR in the post-genomic era"
Juniata College
Huntingdon, Pennsylvania, November 19, 2002.

(38) "GFT NMR Spectroscopy"
Rutgers University
Piscataway, New Jersey, December 11, 2002.

(39) 'NMR spectroscopy for structural genomics'
Rutgers University
Piscataway, New Jersey, January 22, 2002.

(40) 'GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information'
Keystone Symposium, Frontiers in Structural Biology
Taos, New Mexico, February 8, 2003.

(41) 'NMR methodology for structural genomics'
Rutgers University
Piscataway, New Jersey, February 13, 2003.

(42) 'NMR in the postgenomic era'
Roswell Park Cancer Institute
Buffalo, New York, February 24, 2003.

(43) 'NMR-based Structural Genomics: New Methods and Perspectives'
New York State Proteomics Symposium
Syracuse, New York, March 17, 2003.

(44) 'GFT NMR: rapid and precise NMR data collection'
Varian Users Meeting
Savannah, Georgia, March 29, 2003.

(45) 'Implementation of GFT NMR experiments'
Bruker Users Meeting
Savannah, Georgia, March 29, 2003.

(46) 'GFT NMR Spectroscopy: Theory and Applications'
44th Experimental NMR Conference
Savannah, Georgia, April 4, 2003.

(47) 'NMR Methodology for Structural Genomics'
Buffalo Excellence in Biological Sciences Seminar Series
Buffalo, New York, May 15, 2003.

(48) ''RD and GFT NMR: new NMR methods for rapid protein structure determination'
Middle Atlantic Regional Meeting of the ACS
Princeton, New York, June 11, 2003.

(49) 'NMR for structural biology and metabolic profiling'
Northeastern Regional Meeting of the ACS
Saratoga Springs, New York, June 18, 2003.

(50) 'Profiling Yeast Metabolism by NMR'
National meeting of the American Chemical Society
New York, New York, September 12, 2003.

(51) 'NMR at UB: Structural Genomics and Metabolic Flux Profiling'
Seminar Series
Roswell Park Cancer Institute
Buffalo, New York, February 20, 2004

(52) 'GFT NMR News'
45th Experimental NMR Conference,
Asilomar, California, April 23, 2004.

(53) "NMR Methods Enabling Rapid Data Collection"
EMSL meeting 2004, Pacific Northwest National Laboratories
Redland, Washington, June 16, 2004.

(54) 'Rapid Acquisition of Multidimensional NMR data'
Gordon Research Conference in Stereochemistry
Salve Regina University
Newport, Rhode Island, June 21, 2004

(55) 'GFT NMR – Toward HTP NMR Structure Determination'
Departmental Lecture Series, Chemistry Department
University of Rochester
Rochester, New York, October 27, 2004

(56) 'Rapid Sampling of NMR Data'
International Conference on Structural Genomics
Washington, DC, November 20, 2004

(57) 'Fast Acquisition of Multidimensional NMR Data: Implications for Structural Genomics'
Departmental Lecture Series, Chemistry Department
Fayetteville, Arkansas, March 28, 2005

(58) 'GFT NMR – Rapid Protein NMR Data Collection for Structural Genomics'
Departmental Lecture Series, Chemistry Department
Seattle, Washington, March 30, 2005

(59) 'New Methodology for PSI-2: GFT NOESY and G2FT NMR
NESG NMR Division Workshop
Buffalo, New York, May 24, 2005

(60) 'Where do we stand on HTP NMR Structure Determination'
NESG Retreat
Princeton, New Jersey, June 21, 2005

(61) 'GFT NMR-based Structural Genomics'
Keystone Symposia, Frontiers in Structural Biology
Keystone, Colorado, January 31, 2006

2. International

(1) "Synergy of ^{13}C-labeling of Amino Acids and Metabolic Flux Balancing- a Novel Approach to Support Process Design in Biotechnology"
International Conference on Magnetic Resonance in Biological Systems,
Tokyo, Japan, August 27, 1998.

(2) "Exploration of Central Carbon Metabolism using Biosynthetic Fractional ^{13}C-Labeling and Two-dimensional NMR Spectroscopy"
Metabolic Engineering Conference II
Elmau, Germany, October 16, 1998.

(3) "METAFoR by NMR Analysis for Biotechnology Research"
ETH Zürich
Zürich, Switzerland, September 2, 1999.

(4) "NMR Structure of a Chimeric Hybrid Duplex Formed During Initiation of HIV-1 Reverse Transcription"
4th Annual Workshop "Structure-Function Analysis of Drug Resistant HIV-RT"
Rome, Italy, November 12, 1999.

(5) "Neuere Erkenntnisse über lebende Systeme mittels NMR Spectroskopie"
Universität Düsseldorf
Düsseldorf, Germany, January 10, 2000.

(6) "^{13}C-labeling Experiments in Support of Biotechnology Research"
Swiss National Science Foundation Symposium: The Swiss Priority Project Biotechnology,
ETH Zürich,
Zürich, Switzerland, March 23, 2000.

(7) "Towards Structural Biology in Supercooled Water. Implications for the Structure of FluA"
Technische Universität München
München, Germany, March 29, 2000.

(8) "Kernresonanzspektroskopie - faszinierende neue Möglichkeiten zur Ergründung biomolekularer Vorgänge"
Universität Siegen
Siegen, Germany, July 4, 2000.

(9) "Reduced Dimensionality NMR for Structural Genomics"
Protein Engineering Network of Centers of Excellence
Toronto, Canada, October 19, 2000.

(10) "Rapid NMR assignment of Proteins for High-throughput Structure Determination"
1st International Conference on Structural Genomics (ICSG) 2000
Yokohama, Japan, November 3, 2000.

(11) "Structural Biology in Supercooled Water"
NMR in Molecular Biology, European Science Foundation (ESF)
Karrebaeksminde, Denmark, June 11, 2001.

(12) "Structural Genomics by NMR"
Institute of Biotechnology
Vilnius, Lithuania, October 5, 2001.

(13) "Structural Biology in Supercooled Water"
Institute of Biotechnology
Vilnius, Lithuania, August 6, 2002.

(14) "Structural Biology in Supercooled Water"
20th International Conference on MR in Biological Systems
Toronto, Canada, August 27, 2002.

(15) "Flux Information from NMR Data"
FEBS Course "Advanced Technologies For Metabolic Engineering in Biotechnology and Medicine"
Lisbon, Portugal, September 7-14, 2002.

(16) "NMR for Metabolic Profiling: New Insights"
FEBS Course "Advanced Technologies For Metabolic Engineering in Biotechnology and Medicine"
Lisbon, Portugal, September 7-14, 2002.

(17) 'GFT NMR'
Biochemistry Department, University of Toronto
Toronto, Canada, February 27, 2003.

(18) 'NMR in high-throughput: Structural Genomics and Metabolic Flux Profiling'
AstraZeneca Biotechnology Seminar Series
Mississauga, Canada, February 27, 2003.

(19) 'Strukturelle Genomik: Semiempirische Lösung des Proteinfaltungsproblems?'
Seminar Series of the 'Bayreuther Zentrum für Molekulare Biowissenscahften'
Bayreuth, Germany, April 25, 2003.

(20) 'GFT NMR spectroscopy: Rethinking Multidimensional Data Acquisition'
16th International Conference on NMR Spectroscopy
Cambridge, UK, July 1, 2003.

(21) 'GFT NMR for rapid NMR data acquisition'
Jahrestagung der Fachgruppe Resonanzspektroskopie (GDCh)
Leipzig, Germany, October 2, 2003.

(22) 'Structural Genomics by NMR: Novel Methods and Insights'
3rd NCCR Symposium on New Trends in Structural Biology
Switzerland, Zürich, November 15, 2003.

(23) 'GFT NMR, Progress for Rapid NMR Data Collection'
21st International Conference on Magnetic Resonance in Biological Systems
Hyderabad, India, January 2005

(24) 'GFT NMR Based Protein Structure Determination in High-Throughput'
Keystone Symposium, Frontiers in Structural Biology
Banff, Canada, February 1, 2005

(25) 'Strukturelle Genomik: Revolution in Grundlagenforschung und Medikamentenentwicklung'
Virtoweb, Support of IT for Biotech
Bochum, Germany, February 18, 2005

(26) 'Protocol for NMR-based Structural Proteomics'
HUPO 4th Annual World Congress
Munich, Germany, August 29, 2005

(27) 'Studies of the M.HhaI – DNA system'
Institute for Biotechnology
Vilnius, Lithuania, August 18, 2005

(28) 'NMR-based Structural Genomics'
Chemistry Department Lecture Series
Reykjavik, Iceland, October 7, 2005

(29) 'Methodology for NMR-based Structural Genomics'
NMR department, Max-Planck Institute for Biophysical Chemistry
Göttingen, Germany, October 24, 2005

(30) 'GFT Projection NMR Spectroscopy'
University of Halle Lecture Series
Halle, Germany, October 27, 2005

(31) 'NMR for Structural Genomics'
SCAI of the Fraunhofer Gesellschaft
Bonn, Germany, November 9, 2005

(32) 'GFT NMR for structural and dynamic studies of proteins in solution '
7th Igler NMR-symposium
Obergurgl, Austria, February, 2006

EXHIBIT 2

Prof. Dr. Thomas Szyperski
The State University of New York at Buffalo
Department of Chemistry, 816 NSC
Buffalo, NY 14260, USA

Publications

1. Szyperski, T. and Schwerdtfeger, P. (1989) On the Stability of Trioxo(η^5-Cyclopentadienyl) Compounds of Manganese, Technetium and Rhenium: An *ab initio* SCF Study. *Angew. Chem. Int. Ed. Engl.* **28**, 1228–1231.
2. Neri, D., Szyperski, T., Otting, G., Senn, H. and Wüthrich, K. (1989) Stereospecific Nuclear Magnetic Resonance Assignments of the Methyl Groups of Valine and Leucine in the DNA-Binding Domain of the 434 Repressor by Biosynthetically Directed Fractional ^{13}C Labeling. *Biochemistry* **28**, 7510–7516.
3. Szyperski, T., Neri, D., Leiting, B., Otting, G. and Wüthrich, K. (1992) Support of ^{1}H NMR Assignments In Proteins by Biosynthetically Directed Fractional ^{13}C-labeling. *J. Biomol. NMR* **2**, 323–334.
4. Szyperski, T., Güntert, P., Otting, G. and Wüthrich, K. (1992) Determination of Scalar Coupling Constants by Inverse Fourier Transformation of In-Phase Multiplets. *J. Magn. Reson.* **99**, 552–560.
5. Szyperski, T, Güntert. P., Stone, S. R. and Wüthrich, K. (1992) Nuclear Magnetic Resonance Solution Structure of Hirudin(1-51) and Comparison with Corresponding Three-dimensional Structures Determined Using the Complete 65-Residue Hirudin Polypeptide Chain. *J. Mol. Biol.* **228**, 1193–1205.
6. Szyperski, T., Güntert, P., Stone, S. R., Tulinsky, A., Bode, W., Huber, R. and Wüthrich, K. (1992) Impact of Protein-Protein Contacts on the Conformation of Thrombin-bound Hirudin Studied by Comparison with the Nuclear Magnetic Resonance Solution Structure of Hirudin(1-51). *J. Mol. Biol.* **228**, 1206–1211.
7. Wüthrich, K., Szyperski, T., Leiting, B. and Otting, G. (1992) Biosynthetic Pathways of the Common Proteinogenic Amino Acids Investigated by Fractional ^{13}C Labeling and NMR Spectroscopy. In: Frontiers and New Horizons in Amino Acid research (K. Takai, Ed.), Elsevier, Amsterdam, pp 41–48.
8. Szyperski, T., Wider, G., Bushweller, J. H. and Wüthrich, K. (1993) 3D ^{13}C-^{15}N Heteronuclear Two-spin Coherence Spectroscopy for Polypeptide Backbone Assignments in ^{13}C-^{15}N-double Labeled Proteins. *J. Biomol. NMR* **3**, 127–132.
9. Szyperski, T., Luginbühl, P., Otting, G., Güntert, P. and Wüthrich, K. (1993) Protein Dynamics studied by Rotating Frame ^{15}N Spin Relaxation Times. *J. Biomol. NMR* **3**, 151–164.
10. Szyperski, T., Wider, G., Bushweller, J. H. and Wüthrich, K. (1993) Reduced Dimensionality in Triple Resonance Experiments. *J. Am. Chem. Soc.* **115**, 9307–9308.
11. Szyperski, T., Scheek, S., Johansson, J., Assmann, G., Seedorf, U. and Wüthrich, K. (1993) NMR determination of the Secondary Structure and the Three-dimensional Polypeptide Backbone Fold of the Human Sterol Carrier Protein 2. *FEBS Lett.* **335**, 18–26.
12. Johansson, J., Szyperski, T., Curstedt, T. and Wüthrich, K. (1994) The NMR Structure of the Pulmonary Surfactant-Associated Polypeptide SP-C in an Apolar Solvent Contains a Valyl-Rich α-Helix. *Biochemistry* **33**, 6015–6023.

13. Szyperski, T., Antuch, W., Schick, M., Betz, A., Stone, S. R. and Wüthrich, K. (1994) Transient Hydrogen Bonds Identified on the Surface of the NMR Solution Structure of Hirudin. *Biochemistry* **33**, 9303–9310.
14. Ottiger, M., Szyperski, T., Luginbühl, P., Ortenzi, C., Luporini, P., Bradshaw, R. A. and Wüthrich, K. (1994) The NMR Solution Structure of the Pheromone Er-2 From the Ciliated Protozoan *Euplotes raikovi*. *Protein Science* **3**, 1515–1526.
15. Szyperski, T., Pellecchia, M. and Wüthrich, K. (1994) 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN, a Projected 4D NMR Experiment for the Sequential Correlation of Polypeptide $^{1}H^{\alpha/\beta}$, $^{13}C^{\alpha/\beta}$ and Backbone ^{15}N and $^{1}H^{N}$ Chemical Shifts. *J. Magn. Reson. B* **105**, 188–191.
16. Szyperski, T., Pellecchia, M., Wall, D., Georgopoulos, C. and Wüthrich, K. (1994) NMR Structure Determination of the *Escherichia coli* DnaJ Molecular Chaperone: Secondary Structure and Backbone Fold of the N-terminal Region 2-108 Comprising the Highly Conserved J-Domain. *Proc. Natl. Acad. Sci. USA* **91**, 11343–11347.
17. Smith, P. E., van Schaik, R. C., Szyperski, T., Wüthrich, K. and van Gunsteren, W. F. (1995) Internal Mobility of the Basic Pancreatic Trypsin Inhibitor in Solution: A Comparison of NMR Spin Relaxation Measurements and Molecular Dynamics Simulations. *J. Mol. Biol.* **246**, 356–365.
18. Johansson, J., Szyperski, T. and Wüthrich, K. (1995) Pulmonary Surfactant-Associated Polypeptide SP-C in Lipid Micelles: CD Studies of Intact SP-C and NMR Secondary Structure of Depalmitoyl-SP-C(1–17). *FEBS Lett.* **362**, 261–265.
19. Szyperski, T., Braun, D., Fernández, C., Bartels, C. and Wüthrich, K. (1995) A Novel Reduced-Dimensionality Triple Resonance Experiment for Efficient Polypeptide Backbone Assignment, 3D $\underline{CO}$HNN$\underline{CA}$. *J. Magn. Reson. B* **108**, 197–203.
20. Szyperski, T. (1995) Biosynthetically Directed Fractional ^{13}C-labeling of Proteinogenic Amino Acids. An Efficient Analytical Tool to Investigate Intermediary Metabolism. *Eur. J. Biochem.* **232**, 433–448.
21. Luginbühl, P., Szyperski, T. and Wüthrich, K. (1995) Statistical Basis for the Use of $^{13}C^{\alpha}$ Chemical Shifts in Protein Structure Determination. *J. Magn. Reson. B* **109**, 229–233.
22. Zerbe, O., Szyperski, T., Ottiger, M. and Wüthrich, K. (1996) 3D ^{1}H-TOCSY-relayed ct-[^{13}C,^{1}H]-HMQC for Aromatic Spin System Identification in Uniformly ^{13}C Labeled Proteins. *J. Biomol. NMR* **7**, 99–106.
23. Pellecchia, M., Szyperski, T., Wall, D., Georgopoulos, C. and Wüthrich, K. (1996) NMR Structure of the J-domain and the Gly/Phe-rich Region of the *Escherichia Coli*. Dnaj Chaperone. *J. Mol. Biol.* **260**, 236–250.
24. Szyperski, T., Braun, D., Banecki, B. and Wüthrich, K. (1996) Useful Information from Axial Peak Magnetization in Projected NMR Experiments. *J. Am. Chem. Soc.* **118**, 8147–8148.
25. Szyperski, T., Bailey, J. E. and Wüthrich, K. (1996) Detecting and Dissecting Metabolic Fluxes Using Biosynthetic Fractional ^{13}C-labeling and Two-dimensional NMR Spectroscopy. *Trends in Biotechnology* **14**, 453–459.
26. Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mösinger, E. and Wüthrich, K. (1997) NMR Solution Structure of the Pathogenesis-Related Protein P14a. *J. Mol. Biol.* **266**, 576–593.
27. Pellecchia, M., Iwai, H., Szyperski, T. and Wüthrich, K. (1997) The 2D NMR Experiments H(C)CO_2 and H$\underline{CCO_2}$ for Assignment and pH Titration of

Carboxylate Groups in Uniformly $^{15}N/^{13}C$-Labeled Proteins. *J. Magn. Reson.* **124**, 274–278.

28. Sauer, U., Hatzimanikatis, V., Bailey, J. E., Hochuli, M., *Szyperski, T. and Wüthrich, K. (1997) Metabolic Fluxes in Riboflavin-producing *Bacillus subtilis*. *Nature Biotechnol.* **15**, 448–452.
29. Szyperski, T., Fernández, C. and Wüthrich, K. (1997) Two-dimensional ct-HC(C)H-COSY for Resonance Assignments of Smaller ^{13}C-labeled Biomolecules. *J. Magn. Reson.* **128**, 228–232.
30. Szyperski, T., Ono, A., Fernández, C., Iwai, H., Tate, S., Wüthrich, K. and Kainosho, M. (1997) Measurement of $^{3}J_{C2'P}$ Scalar Couplings in a 17 kDa Protein Complex with $^{13}C,^{15}N$-Labeled DNA Distinguishes the B_I and B_{II} Phosphate Conformations of the DNA. *J. Am. Chem. Soc.* **119**, 9901–9902.
31. Klimasauskas, S., *Szyperski, T., Serva, S. and Wüthrich, K. (1998) Dynamic Modes of the Flipped-out Cytosine during *Hha*I Methyltransferase-DNA Interactions in Solution. *EMBO J.* **17**, 371–324.
32. Szyperski, T., Fernández, C., Ono, A., Kainosho, M. and Wüthrich, K. (1998) Measurement of Deoxyribose $^{3}J_{HH}$ Scalar Couplings Reveals Protein-Binding Induced Changes in the Sugar Puckers of the DNA. *J. Am. Chem. Soc.* **120**, 821–822.
33. Szyperski, T., Fernández, C., Mummenthaler, C. and Wüthrich, K. (1998) Structure Comparison of Human Glioma Pathogenesis-Related Protein GliPR and the Plant Pathogenesis-related Protein P14a Indicates a Functional Link between the Human Immune System and a Plant Defense System. *Proc. Natl. Acad. Sci. USA* **95**, 2262–2266.
34. Szyperski, T., Banecki, B., Braun, D. and Glaser, R. W. (1998) Sequential Assignment of Medium-sized $^{15}N/^{13}C$-labeled Proteins with Projected 4D Triple Resonance NMR Experiments. *J. Biomol. NMR* **11**, 387–405.
35. Fernández, C., Szyperski, T., Ono, A., Iwai, H., Tate, S.-I., Kainosho, M. and Wüthrich, K. (1998) NMR with $^{13}C,^{15}N$-doubly-labeled DNA: the *Antennapedia* Homeodomain Complex with a 14mer DNA Duplex. *J. Biomol. NMR* **12**, 25–37.
36. Szyperski, T. (1998) ^{13}C-NMR, MS and Metabolic flux Balancing in Biotechnology Research. *Q. Rev. Biophys.* **31**, 41–106.
37. Weber, F. E., Dyer, J. H., López Garcia, F., Szyperski, T., Wüthrich, K. and Hauser, H. (1998) In Pre-sterol Carrier Protein 2 (SCP2) in Solution the Leader Peptide 1-20 is Flexibly Disordered and the Residues 21-143 Adopt the Same Globular Fold as in Mature SCP2. *Cell. Mol. Life Sci.* **54**, 751–759.
38. Szyperski, T., Vandenbussche, G., Curstedt, T., Ruyschaert, J.-M., Wüthrich, K. and Johansson, J. (1998) Monomeric α-helical Pulmonary Surfactant-associated Polypeptide C Dissolved in a Mixed Organic Solvent Transforms Into Insoluble β-sheet Aggregates. *Protein Sci.* **7**, 2533–2540.
39. Pervushin, K., Ono, A., Fernandez, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998) NMR Scalar Couplings Across Watson-Crick Base Pair Hydrogen Bonds in DNA Observed by Transverse Relaxation-Optimized Spectroscopy. *Proc. Natl. Acad. Sci. USA* **95**, 14147–14151.
40. Fiaux, J., Andersson, C. I. J., Holmberg, N., Bülow, L., Kallio, P. T., Szyperski, T., Bailey, J. E. and Wüthrich, K. (1999) ^{13}C NMR Flux Ratio Analysis of *Escherichia coli* Central Carbon Metabolism in Microaerobic Bioprocesses. *J. Am. Chem. Soc.* **121**, 1407–1408.
41. Szyperski, T., Götte, M., Billeter, M., Perola, E., Cellai, L., Heumann, H. and Wüthrich, K. (1999). NMR Structure of r(gcacuggc)•r(gcca)d(CTGC), a

Chimeric Hybrid Duplex Comprising the tRNA-DNA Junction Formed During the Initiation of HIV-1 Reverse Transcription. *J. Biomol. NMR* **13**, 343–355.

42. Szyperski, T., Glaser, R. W., Hochuli, M., Fiaux, J., Sauer, U., Bailey, J. E. and Wüthrich, K.(1999) Bioreaction Network Topology and Metabolic Flux Ratio Analysis by Fractional ^{13}C-Labeling and Two-dimensional NMR Spectroscopy. *Metabolic. Eng.* **1**, 189–197.
43. Hochuli, M., Patzelt, H., Österhelt, D., Wüthrich, K. and *Szyperski, T. (1999) Amino Acid Metabolism in the Halophilic Archaeon *Haloarcula hispanica*. *J. Bacteriol.* **181**, 3226–3237.
44. Szyperski, T., Fernandez, C., Ono, A., Wüthrich, K. and Kainosho, M. (1999) The {^{31}P}-Spin-echo-difference Constant-time [^{13}C,^{1}H]-HMQC Experiment for Simultaneous Determination of $^{3}J_{H3'P}$ and $^{3}J_{C4'P}$ in Nucleic Acids and their Protein Complexes. *J. Magn. Reson.* **140**, 491–494.
45. Fernandez, C., Szyperski, T., Billeter, M., Ono, A., Iwai, H., Kainosho, M. and Wüthrich, K. (1999) Conformational Changes of the BS2 Operator DNA upon Complex Formation with the *Antennapedia* Homeodomain Studied by NMR with ^{13}C/^{15}N-labeled DNA. *J. Mol. Biol.* **292**, 609–617.
46. Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K. and Bailey, J. E. (1999) Metabolic Flux Ratio (METAFoR) Analysis of Genetic and Environmental Modulations of *Escherichia coli* Central Carbon Metabolism. *J. Bacteriol.* **181**, 6679–6688.
47. Lopez, F., Szyperski, T., Choinowski, T., Dyer, J. H., Hauser, H. and Wüthrich, K. (2000) NMR Structure of the Sterol Carrier Protein-2: Implications for the Biological Role. *J. Mol. Biol.* **295**, 595–603.
48. Sauer, U., Szyperski, T and Bailey, J. E. (2000) Future Trends in Complex Reaction Studies. In: *NMR in Microbiology: Theory and Application* (J.-N. Barbotin and J.-C. Portais, Eds.), Horizon Scientific Press, Norfolk.
49. Skalicky, J. J. and Szyperski, T. (2000) Two-dimensional NMR. In: *NMR in Microbiology: Theory and Application* (J.-N. Barbotin and J.-C. Portais, Eds.), Horizon Scientific Press, Norfolk.
50. Hochuli, M., Szyperski, T. and Wüthrich, K. (2000) Deuterium Isotope Effects on the Central Carbon Metabolism of *Escherichia coli* cells grown on a D_2O-containing Minimal Medium. *J. Biomol. NMR* **17**, 33–42.
51. Skalicky, J. J., Sukumaran, D. K., Mills, J. L. and *Szyperski, T. (2000) Toward Structural Biology in Supercooled Water. *J. Am. Chem. Soc.* **122**, 3230–3231.
52. Montelione, G. T., Zheng, D., Huang, Y., Gunsalus, K. C. and Szyperski, T. (2000) Protein NMR Spectroscopy for Structural Genomics. *Nature Struc. Biol.* **7**, 982–984.
53. Frey, A. D., Fiaux, J., Szyperski, T., Bailey, J. E., Wüthrich, K. and Kallio, P. T. (2001) Dissection of the Central Carbon Metabolism of Hemoglobin-Expressing *Escherichia Coli* by ^{13}C NMR Flux Distribution Analysis in Microaerobic Bioprocesses. *Appl. Environ. Microbiol.* **67**, 680–687.
54. Skalicky, J. J., Mills, J. L., Sharma, S. and *Szyperski, T. (2001) Aromatic Ring-flipping in Supercooled Water: Implications for NMR-based Structural Biology of Proteins. *J. Am. Chem. Soc.* **123**, 388–397.
55. Canonaco, F., Hess, T. A., Wang, T., Szyperski, T. and Sauer, U. (2001) Metabolic Flux Response to Phosphoglucose Isomerase Knock-out in *Escherichia Coli*. *FEMS Microbiol. Lett.* **204**, 247–252.
56. Maaheimo, H., Fiaux, J., Cakar, Z. P., Bailey, J. E., Sauer, U. and *Szyperski, T. (2001) Central Carbon Metabolism of *Saccharomyces cerevisiae* Explored

by Biosynthetic Fractional ^{13}C Labeling of Common Amino Acids. *Eur. J. Biochem.* **268**, 2464–2479.

57. Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wüthrich, K., Bailey, J. E. and Sauer, U. (2001) Metabolic Flux Response to Pyruvate Kinase Knockout in *Escherichia Coli*. *J. Bacteriol.* **184**, 152–164.
58. Dauner, M., Sonderegger, M., Hochuli, M., Szyperski, T., Wüthrich, K., Hohmann, H. P., Sauer, U. (2002) Metabolic Fluxes in Riboflavin-Producing *Bacillus Subtilis* During Growth on Two-carbon Substrate Mixtures. *Appl. Environ. Microbiol.* **68**, 1760–1771.
59. Monleon, D., Colson, K., Moseley, H. N. B., Anklin, C., Oswald, R., Szyperski, T. and Montelione, G. T. (2002) Rapid Analysis of Protein Backbone Resonance Assignments using Cryogenic Probes, a Distributed Linux-based Computing Architecture, and an Integrated Set of Spectral Analysis Tools. *J. Struc. Func. Genomics* **2**, 93-101.
60. *Szyperski, T., Yeh, D. C., Sukumaran, D. K., Moseley, H. N. B. and Montelione, G. T. (2002) Reduced-dimensionality NMR spectroscopy for High-Throughput Resonance Assignment. *Proc. Natl. Acad. Sci. USA* **99**, 8009–8014.
61. Szymczyna, B. R., Pineda-Lucena, A., Mills, J. L., Szyperski, T. and Arrowsmith, C. (2002) ^{1}H, ^{13}C and ^{15}N resonance Assignments and Secondary Structure of the RNA-Binding PWI Domain from SRm160 using Reduced Dimensionality NMR Spectroscopy. *J. Biomol. NMR* **22**, 299-300.
62. Mills, J. L. and *Szyperski, T. (2002) Protein Dynamics in Supercooled Water: The Search for Slow Motional Modes. *J. Biomol. NMR* **23**, 63–67.
63. Szyperski, T. (2002) Strukturelle Genomik. *Nachrichten aus der Chemie* **50**, 1128–1131.
64. Xia, Y., Arrowsmith, C. H. and *Szyperski, T. (2002) Novel Projected 4D Triple Resonance Experiments for Polypeptide Chemical Shift Assignment. *J. Biomol. NMR* **24**, 41–50.
65. Gong, B., Zeng, H., Zhu, J., Yuan, L., Han, Y., Cheng, S., Furukawa, M., Parra, R. D., Kovalevsky, A. Y., Mills, J. L., Skrzypczak-Jankun, E., Martinovic, S., Smith, R. D., Zheng, C., Szyperski, T. and Zeng, X. C. (2002) Creating Nanocavities of Tunable Sizes: Hollow Helices. *Proc. Natl. Acad. Sci. USA* **99**, 11583–11588.
66. Fiaux, J., Cakar, Z. P., Sonderegger, M., Wüthrich, K., *Szyperski, T. and Sauer, U. (2003) Metabolic Flux Profiling of the Yeasts *Saccharomyces cerevisiae* and *Pichia stipitis*. *Eucaryotic Cell* **2**, 170–180.
67. Kim, S. and *Szyperski, T. (2003) GFT NMR, a New Approach to Rapidly Obtain Precise High Dimensional NMR Spectral Information. *J. Am. Chem. Soc.* **125**, 1385–1393.
68. Daujotyte, D., Vilkaitis, G., Manelyt, L., Skalicky, J., *Szyperski, T. and Klimasauskas, S. (2003) Solubility Engineering *HhaI* Methyltransferase for NMR Structural Studies. *Protein Eng.* **16**, 295-301.
69. Liu, G., Mills, J. L., Hess, T. A., Kim, S., Skalicky, J. J., Sukumaran, D. K., Kupce, E., Skerra, A., *Szyperski, T. (2003) Resonance Assignments for the 21 kDa Engineered Fluorescein-binding Lipocalin FluA. *J. Biomol. NMR* **27**, 187–188.
70. Aramini, J. M., Mills, J. L., Xiao, R., Acton, T. B., Wu, M. J., Szyperski, T. and Montelione, G. T. (2003) Resonance Assignments for the Hypothetical Protein yggU from *Escherichia coli*. *J. Biomol. NMR* **27**, 285–286.

71. Monleon, D., Chiang, Y., Aramini, J., Swapna, G.V.T., Palacios, D., Gunsalus, K.C., Kim, S., Szyperski, T. and Montelione, G. T. (2004) Resonance Assignments for the 21 kDa *Caenorhabditis elegans* Homologue of 'Brain-specific' Protein. *J. Biomol. NMR* **28**, 91–92.
72. Kim, S. and *Szyperski, T. (2004) GFT Triple Resonance NMR Experiments for Polypeptide Chemical Shift Assignment. *J. Biomol. NMR* **28**, 117–130.
73. Xu, D., Liu, G., Rong, X., Acton, T., Goldsmith-Fischman, S., Honig, B., Montelione, G. T. and *Szyperski, T. (2004) NMR Structure of the Hypothetical Protein AQ-1857 Encoded by the YI57 Gene from *Aquifex aeolicus* Reveals a Novel Protein Fold. *Proteins* **54**, 794–796.
74. Liu, G., Sukumaran, D. K., Xu, D., Chiang, Y., Acton, T., Goldsmith-Fischman, S., Honig, B., Montelione, G. T. and *Szyperski, T. (2004) NMR Structure of the Hypothetical Protein NMA1147 from *Neisserica meningitidis* Reveals a Distinct 5-helix Bundle. *Proteins* **55**, 756–758.
75. Herve du Penhoat, C., Atreya, H. S., Shen, Y., Liu, G., Acton, T. B., Li, Z., Murray, D., Montelione, G. T. and *Szyperski, T. (2004) The NMR Solution Structure of the 30S Ribosomal Protein S27e Encoded in the Gene RS27_ARCFU of *Archaeoglobus fulgidis* Reveals a Novel Protein Fold. *Protein Sci.* **13**, 1407–1416.
76. Sola, A., Maaheimo, H., Ylonen, K., Ferrer, P. and *Szyperski, T. (2004) Amino Acid Biosynthesis and Metabolic Profiling of *Pichia pastoris*. *Eur. J. Biochem*. **271**, 2462–2470.
77. Zamboni, N., Maaheimo, H., Szyperski, T., Hohmann, H.-P. and Sauer, U. (2004) The Phosphoenolpyruvate Carboxykinase also Catalyzes C3 Carboxylation at the Interface of Glycolysis and the TCA Cycle of *Bacillus subtilis*. *Metabolic Eng*. **6**, 277–284.
78. Shen, Y., Atreya, H. S., Xiao, R., Acton, T. B., Shastry, R., Ma, L., Montelione, G. T. and *Szyperski, T. (2004) Resonance Assignment for the 18 kDa Protein CC1736 from *Caulobacter crescentus*, *J. Biomol. NMR* **29**, 549–550.
79. Moseley, H. N. B., Riaz, N., Aramini, J. M., Szyperski, T. and Montelione, G. T. (2004) A Generalized Approach to Automated NMR Peak List Editing: Application to Reduced Dimensionality Triple Resonance Spectra. *J. Magn. Reson.* **170**, 263–277.
80. Atreya, H. S. and *Szyperski, T (2004) G-matrix Fourier Transform NMR Spectroscopy for Complete Protein Resonance Assignment. *Proc. Natl. Acad. Sci. USA* **101**, 9642–9647.
81. Yuan, L., Zeng, H., Yamato, K., Sanford, A. R., Feng, W., Atreya, H. S., Sukumaran, D. K., Szyperski, T. and Gong, B. (2004) Helical Aromatic Oligoamides: Reliable, Readily Predictable Folding from the Combination of Rigidified Structural Motifs. *J. Am. Chem. Soc.* **126**, 16528–16537.
82. Atreya, H. S. and *Szyperski, T. (2005) Rapid NMR Data Collection. *Methods Enzymol.* **394**, 78–108.
83. Huang, Y. J., Moseley, H., Baran, M. C., Arrowsmith, C. H., Powers, R., Tejero, R., Szyperski, T. and Montelione, G. T. (2005) An Integrated Platform for Automated Analysis of Protein NMR Structures. *Methods Enzymol.* **394**, 111–140.
84. Shen, Y., Goldsmith-Fischman, Atreya, H. S., Acton, T., Ma, L., Xiao, R., Honig, B., Montelione, G. T. and *Szyperski, T. (2005) NMR Structure of the 18 kDa Protein CC1736 From *Caulobacter crescentus* Identifies a Member of the 'START' Domain Superfamily and Suggests Residues Mediating Substrate Specificity. *Proteins* **58**, 747–750.

85. Szyperski, T. (2005) Principles and Application of Projected Multidimensional NMR Spectroscopy – G-matrix Fourier Transform NMR. In: *Emerging Principles in Biophysics* (J.L.R. Arrondo and A. Alonso, Eds.), Springer Verlag, New York.
86. Liu, G., Li, Z., Chiang, Y., Acton, T., Montelione, G. T., Murray, D. and *Szyperski, T. (2005) High-quality Homology Models Derived From NMR and X-ray Structures of E. coli Proteins YgdK and SufE Suggest That All Members of the YgdK/SufE Protein Family are Enhancers of Cysteine Desulfurases. *Protein Sci.* **14**, 1597–1608.
87. Szyperski, T. (2005) Protein NMR Spectroscopy. In: *Encyclopedia of Molecular Cell Biology and Molecular Medicine*. Wiley-CH, Weinheim.
88. Herve du Penhoat, C., Li, Z., Atreya, H. S., Kin, S., Yee, A., Xiao, R., Murray, D., Arrowsmith, C. H. and *Szyperski, T. (2005) Solution NMR Structure of *Thermotoga maritima* Protein TM1509 Reveals a Zn-metalloprotease-like Tertiary Structure. *J. Struc. Func. Genomics* **6**, 51–62.
89. Atreya, H. S., Eletsky, A. and *Szyperski, T. (2005) Resonance Assignment of Proteins with High Shift Degeneracy Based on 5D Spectral Information Encoded in G^2FT NMR Experiments. *J. Am. Chem. Soc.* **127**, 4554–4555.
90. Yang, S., Atreya, H. S., Liu, G. and *Szyperski, T. (2005) G-matrix Fourier Transform NOESY Based Protocol for High-Quality Protein Structure Determination. *J. Am. Chem. Soc.* **127**, 9085–9099.
91. Liu, G., Aramini, J., Atreya, H. S., Eletsky, A., Xiao, R., Acton, T. A., Ma, L. C., Montelione, G. T. and *Szyperski, T. (2005) GFT NMR Based Resonance Assignment for the 21 kDa Human Protein UFC1. *J. Biomol. NMR* **32**, 261.
92. Liu, G., Shen, Y., Xiao, R., Acton, T. A., Ma, L. C., Joachimiak, A., Montelione, G. T. and *Szyperski, T. (2005) NMR Structure of Protein yqbG Encoded by Gene YQBG_BASCU From *Bacillus subtilis* Reveals a Novel α-Helical Protein Fold. *Proteins*, in press.
93. Pineda-Lucena, A., Ho, C. S., Mao, D. Y., Sheng, Y., Laister, R. C., Muhandiram, R., Lu, Y., Seet, B. T., Katz, S., Szyperski, T., Penn, L. Z. and Arrowsmith, C. H. (2005) A Structure-based Model of the c-Myc/Bin1 Protein Interaction Shows Alternative Splicing of Bin1 and c-Myc Phosphorylation are Key Binding Determinants. *J. Mol. Biol.* **351**, 182–194.
94. Liu, G., Shen, Y., Atreya, H. S., Parish, D., Shao, Y., Sukumaran, D., Xiao, R., Yee, Adelinda, Lemak, A., Bhattacharya, A., Acton, T. A., Arrowsmith, C. H., Montelione, G. T. and *Szyperski, T. (2005) NMR Data Collection and Analysis Protocol for High-throughput Protein Structure Determination. *Proc. Natl. Acad. Sci. USA* **102**, 10487–10492.
95. *Szyperski, T., Mills, J. L., Perl, D. and Balbach, J. (2005) Combined NMR-observation of Cold Denaturation in Supercooled Water and Heat Denaturation Enables Accurate Measurement of ΔC_p of Protein Unfolding. *Eur. Biophys. J.* **21**, 1–4.
96. Eletsky, A., Atreya, H. S., Liu, G. and *Szyperski, T. (2005) Probing Structure and Functional Dynamics of (large) Proteins with Aromatic Rings: L-GFT-TROSY (4,3)D HCCH NMR Spectroscopy. *J. Am. Chem. Soc.* **127**, 14578–14579.

US patent (number 6.831.459): Method of using G-matrix Fourier transformation nuclear magnetic resonance (GFT NMR) spectroscopy for rapid chemical shift assignment and secondary structure determination of proteins

EXHIBIT 3

High-throughput three-dimensional protein structure determination

Udo Heinemann*†, Gerd Illing‡ and Hartmut Oschkinat†§

In the wake of finished genomic sequencing projects, high-throughput analysis techniques are being developed in various fields of functional genomics. Of special interest in this regard is the three-dimensional structure analysis of proteins by X-ray crystallography and NMR spectroscopy, which has been characterized by distinctly low-throughput in the past. A number of recent advances in instrumentation and software are promising to radically change this situation, leaving the production of suitable protein samples as the sole rate-limiting step in structural analyses.

Addresses
*Forschungsgruppe Kristallographie, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany; e-mail: heinemann@mdc-berlin.de
†Institut für Chemie/Kristallographie, Freie Universität Berlin, Takustrasse 6, 14195 Berlin, Germany
‡PSF biotech AG, Heubnerweg 6, D-14059 Berlin, Germany
§Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany

Current Opinion in Biotechnology 2001, 12:348–354

0958-1669/01/$ – see front matter

Abbreviations

GFP	green fluorescent protein
MAD	multiwavelength anomalous diffraction
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
SAD	single-wavelength anomalous diffraction

Introduction

With the draft sequence of the human genome now available [1,2] and the genomic sequences of many other organisms either known or currently under investigation, the attention of biologists is shifting back to studies of gene products. Research activities in this area, aimed at a complete and systematic study of the distribution, modification and interaction of gene products (usually proteins) in defined tissues, are collectively known as functional genomics [3]. Structural genomics, sometimes called structural proteomics, is an important branch of functional genomics. It is concerned with the systematic three-dimensional structure analysis of proteins and promises to yield a comprehensive mechanistic understanding of cell physiology at the molecular level [4•]. In particular, it is hoped that a complete description of all protein domain folds can be achieved within the near future by solving crystal structures or nuclear magnetic resonance (NMR) structures of representative proteins for all known sequence families.

Since 1999, a number of consortia have formed worldwide to pursue the goals of structural genomics (Table 1). Although these initiatives differ in various respects [5–7], they are united in their resolve to develop and utilize techniques for high-throughput three-dimensional structure analysis. It is usually assumed that an order of magnitude in the speed of structure determination must be gained in order to approach the ambitious goals of structural genomics. In view of this, an industry is developing aimed at high-throughput structural genomics and structure-based drug design (Table 2). Some of the companies are devoted to the development of high-throughput methods, whereas others focus more directly on drug design which is, of course, the driving force behind the commercial applications of structural genomics. In this review we shall outline the general requirements for high-throughput protein

Table 1

Academic consortia devoted to high-throughput three-dimensional protein structure analysis.

Country	Coordinating institution	Protein targets
Canada	University of Toronto	Archaeal proteins
France	Institut Pasteur	Proteins from *Mycobacterium tuberculosis* and other pathogens
	Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch	Protein families from eukaryotes including man
	Université Paris-Sud	Yeast proteins
Germany	Max-Delbrück-Centrum, Berlin	Human proteins
Japan	Genomic Sciences Center, RIKEN, Yokohama Institute	Proteins from animals, plants and *Thermus thermophilus* HB8
	Biological Information Research Center (BIRC), Tokyo	Membrane proteins
United Kingdom	University of Oxford	Human and viral proteins
	Daresbury Laboratory	Proteins from human pathogens
United States	Rutgers University	Eukaryotic protein families
	Argonne National Laboratory	Proteins unique to pathogens or unique to eukaryotes
	University of Georgia	*Pyrococcus furiosus*, *Caenorhabditis elegans* and human proteins
	Rockefeller University	Yeast proteins
	Scripps Research Institute	Human and *C. elegans* signal transduction proteins
	UC Berkeley and Lawrence Berkeley National Laboratory	*Mycoplasma* proteome
	Los Alamos National Laboratory	*M. tuberculosis* proteins

Table 2

Industrial applications of structural genomics.

Company	Founded	Techniques	Focus
Plexxikon Inc., San Francisco, CA	2001	X-ray	Complexes of kinases and ligands
Integrative Proteomics, Toronto, Canada	2000	X-ray, NMR	Structure determination service
PSF biotech AG, Berlin, Germany	2000	X-ray, NMR	High-throughput structure determination, drug discovery
Structural GenomiX Inc., San Diego, CA	1999	X-ray	High-throughput crystallization
Syrrx Inc., San Diego, CA	1999	X-ray	High-throughput crystallization, drug design
Astex Technologies Ltd., Cambridge, UK	1999	X-ray	High-throughput crystallization of protein–ligand complexes
GeneFormatics Inc., San Diego, CA	1998	X-ray, NMR	Structure determination
Proteros Biostructures GmbH, Martinsried, Germany	1998	X-ray, NMR	Structure determination
TRIAD Therapeutics, San Diego, CA	1998	NMR	NMR-supported drug discovery
Bio-X-tal, France	1997	X-ray	Structures of G-protein-coupled receptors
Emerald Biostructures Inc., Bainbridge, WA	1997	X-ray	Structure determination
3D-Pharmaceuticals Inc., Exton, PA	1993	X-ray	Structure determination, drug discovery

structure analysis and summarize recent progress in protein production, crystal and NMR structure determination. We shall end by describing some recent achievements concerning structure analysis speed and throughput.

General requirements for high-throughput three-dimensional structure analysis

In the past, three-dimensional protein structure analysis tended to be a slow endeavor owing to the fact that for each protein molecule many preparative and analytical steps had to be performed in sequence, and many of these steps involved tedious trial-and-error searches for optimum conditions. This held true for both X-ray crystallography and NMR spectroscopy, the two analytical techniques currently available for the determination of protein structures at the atomic level. In this respect, protein purification was notorious as each protein required its own set of buffers and purification columns. Similarly, the search for crystallization conditions and the heavy-atom derivatives needed for phasing in X-ray analysis were frequently time-consuming. Obviously, any attempt to achieve high-throughput in three-dimensional protein structure analysis will require efforts to coordinate, standardize and automate the relevant preparative and analytical procedures [8]. At present, many options are available for the analysis of protein structures, starting with the basic choice of either solving a crystal structure or an NMR solution structure and including many systems for recombinant DNA expression and protein purification (Figure 1). Thus, for each structural analysis it is possible to follow several experimental options in parallel and, of course, several protein structures can be determined at the same time if suitable automation is in place (see below). Automation, in turn, requires a degree of standardization, which can be achieved using currently available technology. Finally, it must be pointed out that both X-ray crystallography and NMR spectroscopy require milligram amounts of very pure protein. Some available methods for high-throughput protein purification [9] are in need of substantial upscaling in order to be useful for three-dimensional structure analysis.

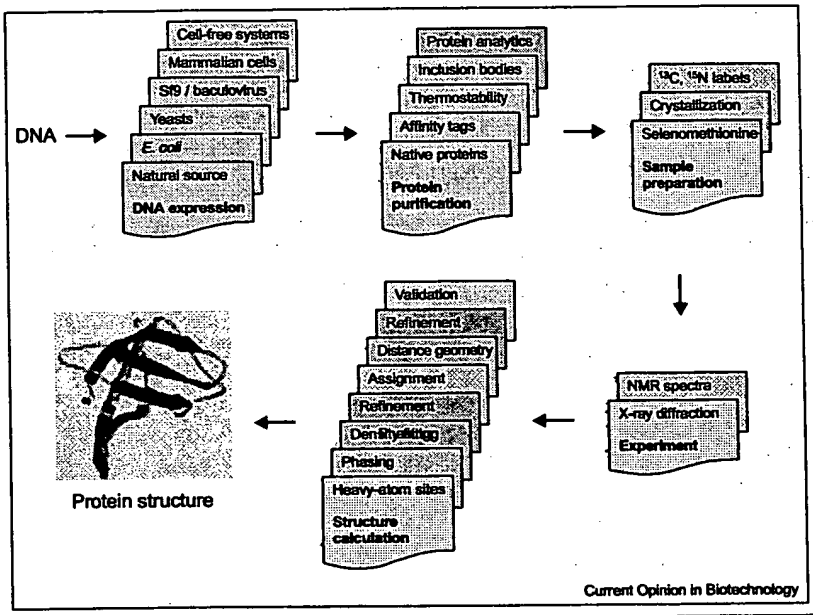
Production of stably folded proteins

A summary of the procedures used for protein production in a high-throughput structural genomics setting has been recently reported [10]. Here, we shall thus limit ourselves to discussing a few selected aspects of protein purification for structural studies. For studies of their three-dimensional structures, proteins must be in their native state and must resist unfolding and degradation for as long as it takes to grow crystals and perform diffraction experiments or to record NMR spectra. Protein solubility in aqueous buffers is often taken as evidence for native-like folding. A more reliable way to assess the folding state of a recombinant protein may be to fuse it to the N terminus of green fluorescent protein (GFP) and to use the fluorescence from GFP as an indicator of correct folding [11•]. The advantage of this method is that it can be applied before purification, thus avoiding preparative effort.

Protein isolation from natural tissues cannot easily be automated and incorporated into a high-throughput structure analysis scheme. Instead, proteins are usually produced by recombinant expression of their genes in various host cells, *Escherichia coli* being the preferred and often easiest choice. It is well known, however, that many proteins cannot be produced in their native state in *E. coli*. For these problem cases, alternative expression systems such as yeast (e.g. *Saccharomyces cerevisiae* and *Pichia pastoris*) and baculovirus-infected insect cells (e.g. Sf9) are in common use. Protein production analysis in mammalian cell culture is expensive and cumbersome and can only be an option of last resort in high-throughput three-dimensional structure analysis. Bacterial cell-free systems present an interesting alternative to protein production in recombinant host cells. High yields of isotope-labeled proteins for NMR structure analysis were recently obtained from a cell-free system based on the S30 fraction of an *E. coli* extract [12•].

In addition to high product yield, protein purification from recombinant host cells offers the advantage that the proteins can be produced in a way that facilitates their subsequent purification. In a high-throughput process it is

Figure 1



The sequence of preparative and analytical steps proceeding from the coding DNA to the three-dimensional protein structure. At each step from DNA expression to structure calculation, procedures or options common to X-ray crystallography and NMR spectroscopy are shown in green, X-ray-specific steps are shown in yellow, and NMR-specific steps in blue. Items shown in the same shade of one basic color mark parallel options; darker shades of a basic color mark items that occur later in time during structure analysis. The term 'protein analytics' summarizes the high-throughput application of mass spectroscopy, gel electrophoresis, infrared spectroscopy and circular dichroism spectroscopy to characterize the purity and the folding state of the proteins.

absolutely essential that purification does not depend on the tedious optimization of column chromatography methods that exploit subtle differences in protein size, charge or hydrophobicity. This can be achieved in a number of ways. Foremost, N- and/or C-terminal tags for specific high-affinity protein binding to suitable resins must be mentioned. Following the now classical introduction of the His_6-tag, which affords binding to a nickel–chelate column and can be reversed by the addition of imidazole [13], a number of further affinity modifications have been described that, in favorable cases, permit one-step purification to homogeneity. As a rule (although not without exception), fusions to proteinogenic tags, such as glutathione-*S* transferase (GST) or maltose-binding protein (MBP), require proteolytic liberation of the target protein before structure analysis. Conversely, with small oligopeptide tags, such as the His- or Strep-tags [14], it may be possible to determine the structure of the tagged protein. The available options as to the choice and combination of tags, their proteolytic removal and handling are too numerous to discuss here. A particularly easy purification is possible for proteins of thermophilic origin synthesized in recombinant *E. coli* strains [15]. Often, these proteins remain soluble after a heat treatment of the crude cell lysate that precipitates a great majority of contaminating host proteins, yielding the protein of interest in nearly pure form. Finally, the high-level expression of foreign genes in *E. coli* often leads to the precipitation of unfolded protein chains and the formation of inclusion bodies. This can be a curse or a blessing depending on whether the protein chains can be refolded into native form or not. Refolding from inclusion bodies can yield almost pure protein, but the refolding success is difficult to assess in the absence of functional protein assays. A way round this dilemma has recently been described in a study that used partial proteolysis to test for stably (re)folded, hydrolysis-resistant protein [16].

Heterologous protein production in recombinant host strains also offers the potential to incorporate labels needed for structure analysis, for instance ^{13}C and/or ^{15}N isotopes for NMR spectroscopy. The labeling of proteins in *E. coli* with selenomethionine to permit multiwavelength anomalous diffraction (MAD) phasing [17] has done away with the trial-and-error searching for heavy-atom derivatives and thus provided one important prerequisite for high-throughput crystal structure analysis. An important advance was recently made when conditions for the incorporation of selenomethionine into proteins in yeast were determined [18].

X-ray diffraction experiments

Typically, it takes several hundred to a thousand individual crystallization experiments to establish the growth conditions needed to produce diffraction-quality protein crystals. If protein three-dimensional structures are to be determined at high-throughput, one quickly reaches numbers of experiments that cannot be handled manually and supervised by eye; protein crystallization begs for automation. At present, several pilot projects worldwide are developing robotic stations for protein crystallization that take care of pipetting, storage, supervision and databank management tasks. An example of such a system with the capacity to handle up to 960,000 experiments simultaneously has recently been described [19].

Achieving high-throughput in protein crystal structure analysis critically depends on the availability of synchrotron

sources. X-rays produced at these facilities are not only much more brilliant than those from laboratory sources, but their energy is also precisely tuneable, thus allowing optimal use to be made of anomalous diffraction for phasing. At high-energy third-generation storage rings, the X-ray light is so brilliant that complete data sets for protein structure analysis can be collected within minutes using fast detectors. Consequently, automation of crystal mounting and handling, along with adequate site management, are required to reduce experimental setup times and make optimal use of the facilities [20]. Several approaches along these lines have been documented [21,22]. Synchrotron users will appreciate efforts to make beamline hardware and software portable and to achieve a degree of standardization in beamline procedures.

Having arrived at the synchrotron beamline, the crystallographer has to decide on their data acquisition and structure solving strategy. We have already mentioned MAD phasing of selenomethionine-substituted proteins [17], which has revolutionized protein crystal structure analysis. There have also been discussions recently about the comparative merit of single-wavelength anomalous diffraction (SAD) phasing, where anomalous X-ray diffraction is observed at one single wavelength and datasets are collected at higher redundancy [23]. Even in cases where selenomethionine incorporation failed, structures might be solved using either MAD or SAD after soaking the protein crystal in solutions of halide salts [24•].

Protein structures from X-ray diffraction data

The analysis of X-ray diffraction data yields novel protein structures through the consecutive stages of locating heavy-atom sites, phasing and computation of electron-density maps, fitting atomic models to these maps, structure refinement and validation (see Figure 1). Recently, there have been significant conceptual and computational advances in all these areas.

Finding sites of anomalous scatterers for MAD or SAD phasing can be difficult in cases where there are weak anomalous signals and a large number of sites. Locating these sites has been facilitated by an automated search procedure, based on the Patterson function, which works at intermediate resolution of the diffraction data [25]. Alternatively, sites may be located by direct methods that exploit statistical relations between structure factors [26,27].

Advances in phasing diffraction data by MAD, SAD, isomorphous replacement and other means have recently been summarized [28]. The MAD technique in various forms [29] is now integrated into a number of semi-automated software packages (e.g. [30]). Often, there is a need to improve initial electron-density maps or to extend them to higher resolution. Near-automated density-modification techniques for this purpose are now available [31]. Finally, there may be special circumstances that allow phase determination in the total absence of heavy-atom markers. One such situation occurs when the X-ray diffraction data extend to truly atomic resolution (~1 Å or better). Using statistical ('direct') methods [26,27] it is then possible to phase the diffraction data and establish the three-dimensional protein structure very rapidly [32]. The challenge here is to extend the resolution range where direct methods can be applied. Another special case for phasing arises when an approximate model of the protein structure to be determined is already known, as is often the case when protein structures are used to guide a pharmacological lead optimization process. The phasing method of molecular replacement used in these cases has been known for quite some time, but new algorithms [33] are proving useful for speeding up the process and solving problem cases.

Another area in which efforts towards automation have met with success is in building atomic models of proteins into electron densities. Using a structural database of protein fragments [34], pattern-matching algorithms [35] or loose-atom partial structures [36•], model building has been greatly accelerated and, to a large extent, automated. These new techniques can be combined with phasing and with structure refinement.

A final and very important issue is the validation of the structure model. High-throughput analysis must not lead to crystal structures of inferior quality. Methods for assessing the internal consistency of protein coordinate sets have been known for some time [37]. More recently, complementary methods for probing the quality of the experimental data and the fit of the atomic model to the data have been introduced [38–40]. The Protein Data Bank [41] is already playing an active role in data and model checking and is expected to continue to do so in the future.

NMR developments towards high-throughput structure determination

The NMR structure determination process in itself consists of a number of different time-consuming steps, independent of sample preparation, which until recently could take up to several months or even years. Recording a data set took a minimum of six to eight weeks, and the assignment of resonance signals, including sidechain signals, required at least three to four weeks. Furthermore, the interpretation of the NOEs (nuclear Overhauser effects) observed in NOESY (NOE spectroscopy)-type spectra could take a couple of months, followed by one or two weeks of structure calculation. To achieve high-throughput, these times have to be significantly reduced. Apart from technical developments to speed up the conventional structure determination process, a reduction of work is achieved in concepts that lead to global folds, trading time for resolution, and utilizing biocomputing techniques.

NMR experiments

The time needed for data acquisition may be reduced by either enhancing the signal-to-noise ratio of spectrometers or by defining structure determination strategies that

require only a minimal data set. In time for the structural genomics age, new probes with supercooled detection and preamplification circuits have been introduced, increasing the signal-to-noise ratio by a factor of three or more and allowing, in principle, for a cut-down in data acquisition times by a factor of nine. In combination with residue-specific deuteration/protonation, data sets may be recorded within two or three weeks [42] that enable the elucidation of reasonably resolved structures or global folds. In combination with homology-search/modeling methods and on the basis of earlier developments [43], the data sets can be further reduced when only the determination of the fold is required [44–46]. The various approaches may utilize experiments that only allow the assignment of backbone atoms and measurements of residual dipolar couplings on partially oriented samples. Other methods use labeling strategies that include extensive systematically selective deuteration of the protein, which leave either only amide protons, methyl groups or other sidechain positions protonated [42,47]. These developments aim to simplify the resonance and NOE assignment process and might be used in combination with homology-search/modeling efforts.

NOE assignment and structure calculation strategies

Until recently, the largest amount of manual interference was devoted to the interpretation of NOESY-type spectra, commonly aided by the semi-automated procedures available in many drawing-board-like software packages [48]. A promising step towards fully automated NOE interpretation is the treatment of a large number or all NOE constraints as ambiguous during structure calculations [49,50•]; however, this approach requires further testing in practice. Work that would previously have taken several months may hopefully soon boil down to preparing a carefully picked peak list, including frequencies and integrals, and refinement cycles of one or two weeks. Intelligent software for generating such peak lists [51] and improvements like routines for estimating line shapes, are hence becoming an important issue. Further developments towards a fully automated structure determination concept will probably have to include back-calculation of two- and three-dimensional NOESY spectra to overcome the shortcomings of peak-picking routines.

Structure validation of NMR structures

On the one hand automation may avoid human errors, but on the other hand artefacts introduced by the automatic procedures may be overlooked. This is a particular problem for NMR, where a quality control system still needs further developments. Hence, the application of software for structure validation is indispensable in a fully automated process, and existing programs that analyse protein coordinates alone (e.g. PROCHECK [37]) need to be routinely applied to such automatically generated NMR structures. Furthermore, it is of critical importance to develop an NMR-specific validation scheme that tests the experimental data. A broad overview of critical parameters is given in [52], together with a compilation of possible pitfalls. A dedicated NMR structure validation program based on a comparison between observed and calculated NOESY spectra has been introduced [53•]; this program calculates an NMR-specific R factor, in analogy to X-ray crystallography.

Recent achievements

Given suitable protein samples, three-dimensional structures can, in principle, be determined quite rapidly using currently available technology. For a small protein, *de novo* protein structure analysis on a third-generation synchrotron storage ring at a station equipped with a large charge-coupled device (CCD) detector can be completed within hours, with the actual diffraction experiment taking just some ten minutes [54••]. The repeated analysis of similar structures (e.g. as part of a pharmacological lead-optimization strategy) may be even faster. Data acquisition for NMR structure analysis is still slower, but might be expected to speed up with the availability of new cryoprobes and new partial labeling techniques [42]. That the speed of data acquisition and structure calculation can indeed translate into high-throughput structure analysis was recently shown in the first report of the achievements of a structural genomics project [55••]. This pilot study on *Methanobacterium thermoautotrophicum* demonstrated the feasibility of high-throughput protein expression, characterization and structure determination. A set of 424 proteins were expressed, and more than 27 structures have been determined to date. In conclusion, the authors cited limitations in hardware accessibility (e.g. synchrotron radiation and NMR time) as a major bottleneck. Also, they pointed out the necessity for automation of the structure determination process.

Conclusions

A number of recent developments in both the fields of X-ray diffraction and NMR spectroscopy hold the promise that fast access to three-dimensional protein structures may be gained once suitable protein crystals or NMR samples are available. The production of these samples may well prove to be the experimental bottleneck.

Acknowledgements

The authors' own work in the field of high-throughput three-dimensional protein structure analysis has been made possible by funding for the Leitprojektverbund 'Proteinstrukturfabrik' through the Federal Ministry of Education and Research.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: The sequence of the human genome. *Science* 2001, 291:1304-1351.
3. Vukmirovic OG, Tilghman SM: Exploring genome space. *Nature* 2000, 405:820-822.

4. • Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S: Structural genomics: beyond the human genome project. *Nat Genet* 1999, 23:151-157.

A clear description of the basic concepts of structural genomics and a discussion of its technical challenges and potential problems.

5. Terwilliger TC: Structural genomics in North America. *Nat Struct Biol* 2000, 7:935-939.

6. Heinemann U: Structural genomics in Europe: slow start, strong finish? *Nat Struct Biol* 2000, 7:940-942.

7. Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzo M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y *et al.*: Structural genomics projects in Japan. *Nat Struct Biol* 2000, 7:943-945.

8. Heinemann U, Frevert J, Hofmann KP, Illing G, Maurer C, Oschkinat H, Saenger W: An integrated approach to structural genomics. *Prog Biophys Mol Biol* 2000, 73:347-362.

9. Lanio T, Jeltsch A, Pingoud A: Automated purification of His_6-tagged proteins allows exhaustive screening of libraries generated by random mutagenesis. *BioTechniques* 2000, 29:338-342.

10. Edwards AM, Arrowsmith CH, Christendat D, Dharamsi A, Friesen JD, Greenblatt JF, Vedadi M: Protein production: feeding the crystallographers and NMR spectroscopists. *Nat Struct Biol* 2000, 7:970-972.

11. • Waldo GS, Standish BM, Berendzen J, Terwilliger TC: Rapid protein folding assay using green fluorescent protein. *Nat Biotechnol* 1999, 17:691-695.

The paper presents evidence that the green fluorescence of *E. coli* cells may indicate the correct folding of proteins fused to the N terminus of GFP. The assay can be used to assess the folding of various protein constructs without requiring protein purification or functional assays.

12. • Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S: Cell-free production and stable isotope labeling of milligram quantities of proteins. *FEBS Lett* 1999, 442:15-19.

A cell-free system based on an *E. coli* S30 extract is used to produce large amounts of ^{13}C- and ^{15}N-labeled protein for NMR analysis.

13. Hochuli E, Bannwarth W, Dobeli H, Gentz R, Stüber D: Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Biotechnology* 1988, 6:1321-1325.

14. Schmidt TGM, Skerra A: One-step affinity purification of bacterially produced proteins by means of the 'Strep-tag' and immobilized recombinant core streptavidin. *J Chromatogr* 1994, 676:337-345.

15. Kim R, Sandler SJ, Goldman S, Yokota H, Clark AJ, Kim SH: Overexpression of archaeal proteins in *Escherichia coli*. *Biotechnol Lett* 1998, 20:207-210.

16. Heiring C, Muller YA: Folding screening assayed by proteolysis: application to various cystine deletion mutants of vascular endothelial growth factor. *Protein Eng* 2001, 14:183-188.

17. Hendrickson WA, Horton JR, LeMaster DM: Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J* 1990, 9:1665-1672.

18. Bushnell DA, Cramer P, Kornberg RD: Selenomethionine incorporation in *Saccharomyces cerevisiae* RNA polymerase II. *Structure* 2001, 9:R11-R14.

19. Mueller U, Nyarsik L, Horn M, Rauth H, Przewieslik T, Saenger W, Lehrach H, Eickhoff H: Development of a technology for automation and miniaturisation of protein crystallisation. *J Biotechnol* 2001, 85:7-14.

20. Abola E, Kuhn P, Earnest T, Stevens RC: Automation of X-ray crystallography. *Nat Struct Biol* 2000, 7:973-977.

21. Cork C, Padmore H, McDermott G, Hung LW, Henderson K, Robinson A, Earnest T: The macromolecular crystallography facility at the advanced light source. *Synchrotron Radiat News* 1998, 11:18-25.

22. Skinner JM, Sweet RM: Integrated software for a macromolecular crystallography synchrotron beamline. *Acta Crystallogr* 1998, 54:718-725.

23. Rice LM, Earnest TN, Brunger AT: Single-wavelength anomalous diffraction phasing revisited. *Acta Crystallogr* 2000, 56:1413-1420.

24. • Dauter Z, Li M, Wlodawer A: Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics. *Acta Crystallogr* 2001, 57:239-249.

Crystals of a pepstatin-insensitive carboxyl proteinase (PCP) were soaked with sodium bromide solution and anomalous diffraction data collected at the peak absorption wavelength. On the basis of these data, the crystal structure could be determined by SAD phasing with a minimum of human intervention.

25. Grosse-Kunstleve RW, Brunger AT: A highly automated heavy-atom search procedure for macromolecular structures. *Acta Crystallogr* 1999, 55:1568-1577.

26. Sheldrick GM: Patterson superposition and *ab initio* phasing. *Methods Enzymol* 1997, 276:628-641.

27. Weeks CM, Miller R: Optimizing shake-and-bake for proteins. *Acta Crystallogr* 1999, 55:492-500.

28. Lamzin VS, Perrakis A: Current state of automated crystallographic data analysis. *Nat Struct Biol* 2000, 7:978-981.

29. Hendrickson WA, Ogata CM: Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol* 1997, 276:494-523.

30. Terwilliger TC, Berendzen J: Automated structure solution for MIR and MAD. *Acta Crystallogr* 1999, 55:849-861.

31. Cowtan KD, Zhang KY: Density modification for macromolecular phase improvement. *Prog Biophys Mol Biol* 1999, 72:245-270.

32. Parisini E, Capozzi F, Lubini P, Lamzin V, Luchinat C, Sheldrick GM: *Ab initio* solution and refinement of two high-potential iron protein structures at atomic resolution. *Acta Crystallogr* 1999, 55:1773-1784.

33. Kissinger CR, Gehlhaar DK, Fogel DB: Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr* 1999, 55:484-491.

34. Diller DJ, Redinbo MR, Pohl E, Hol WG: A database method for automated map interpretation in protein crystallography. *Proteins Struct Funct Genet* 1999, 36:526-541.

35. Holton T, Ioerger TR, Christopher JA, Sacchettini JC: Determining protein structure from electron-density maps using pattern matching. *Acta Crystallogr* 2000, 56:722-734.

36. • Perrakis A, Morris R, Lamzin VS: Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* 1999, 6:458-463.

Using an intermediate free-atom model, phases and electron-density maps were semi-automatically improved and protein models complete with solvent were built into the electron density and refined.

37. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993, 26:283-291.

38. Dodson EJ, Davies GJ, Lamzin VS, Murshudov GN, Wilson KS: Validation tools: can they indicate the information content of macromolecular crystal structures? *Structure* 1998, 6:685-690.

39. van den Akker F, Hol WGJ: Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structures. *Acta Crystallogr* 1999, 55:206-218.

40. Vaguine AA, Richelle J, Wodak SJ: SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr* 1999, 55:191-205.

41. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J: The protein data bank and the challenge of structural genomics. *Nat Struct Biol* 2000, 7:957-959.

42. Medek A, Olejniczak ET, Meadows RP, Fesik SW: An approach for high-throughput structure determination of proteins by NMR spectroscopy. *J Biomol NMR* 2000, 18:229-238.

43. Cornilescu G, Delaglio F, Bax A: Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 1999, 13:289-302.

44. Meiler J, Peti W, Griesinger C: DipoCoup: a versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J Biomol NMR* 2000, 17:283-294.

45. Bowers PM, Strauss CE, Baker D: *De novo* protein structure determination using sparse NMR data. *J Biomol NMR* 2000, 18:311-318.

46. Fowler CA, Tian F, Al-Hashimi HM, Prestegard JH: Rapid determination of protein folds using residual dipolar couplings. *J Mol Biol* 2000, 304:447-460.

47. Goto NK, Kay LE: New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Curr Opin Struct Biol* 2000, 10:585-592.

48. Moseley HN, Montelione GT: Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 1999, 9:635-642.

49. Nilges M, Macias MC, O'Donoghue SI, Oschkinat H: Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J Mol Biol* 1997, 269:408-422.

50. • Xu Y, Jablonsky MJ, Jackson PL, Braun W, Krishna NR: Automatic assignment of NOESY cross peaks and determination of the protein structure of a new world scorpion neurotoxin using NOAH/DIAMOND. *J Magn Reson* 2001, 148:35-46.
The solution NMR structure of a small protein was determined using an automated and a manual approach for resonance assignment and three-dimensional structure calculation. The two sets of structures were shown to be very similar.

51. Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K: Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* 1998, 135:288-297.

52. Doreleijers JF, Rullmann JAC, Kaptein R: Quality assessment of NMR structures: a statistical survey. *J Mol Biol* 1998, 281:149-164.

53. • Gronwald W, Kirchhofer R, Gorler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR: RFAC, a program for automated NMR R-factor estimation. *J Biomol NMR* 2000, 17:137-151.
This paper describes an NMR-specific reliability index that is practical in every-day structure analysis.

54. •• Walsh MA, Dementieva I, Evans G, Sanishvili R, Joachimiak A: Taking MAD to the extreme: ultrafast protein structure determination. *Acta Crystallogr* 1999, 55:1168-1173.
At a third-generation synchrotron source, X-ray diffraction data were used to solve the crystal structure of a selenomethionine-labeled small protein by MAD phasing; the data were collected in less than half an hour. This study gives an impression of the speed of crystal structure analysis that can be achieved.

55. •• Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I *et al.*: Structural proteomics of an archaeon. *Nat Struct Biol* 2000, 7:903-909.
The first published account from a structural genomics project. Focussing on thermostable archaeal proteins facilitated the production of protein samples for NMR spectroscopy and X-ray crystallography.

EXHIBIT 4

Current Topics

Nuclear Magnetic Resonance in the Era of Structural Genomics†

J. H. Prestegard,* H. Valafar, J. Glushka, and F. Tian

Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602

Received January 31, 2001; Revised Manuscript Received June 7, 2001

ABSTRACT: Current interests in structural genomics, and the associated need for high through-put structure determination methods, offer an opportunity to examine new nuclear magnetic resonance (NMR) methodology and the impact this methodology can have on structure determination of proteins. The time required for structure determination by traditional NMR methods is currently long, but improved hardware, automation of analysis, and new sources of data such as residual dipolar couplings promise to change this. Greatly improved efficiency, coupled with an ability to characterize proteins that may not produce crystals suitable for investigation by X-ray diffraction, suggests that NMR will play an important role in structural genomics programs.

The enormous progress that has been made in the sequencing of DNA over the past few years is forcing much of the biochemistry and molecular biology community to reassess its approach to scientific investigation and discovery. Rather than focus on a single system with dogged testing of hypotheses related to mechanisms of action, there is a sense that new insight, or at least new hypotheses, can be generated by taking a broader view and systematically analyzing vast sets of data that have been accumulated in the absence of highly focused objectives. The human genome alone is expected to encode on the order of 30 000 proteins; add to this the genomes of agriculturally important plants and the genomes of pathogenic organisms, and the amount of information to be tapped is staggering (*1*). Most fruits of this new mode of investigation will not come, however, without additional effort, both in analysis and in information gathering. It is estimated that current analysis techniques can assign a function to no more than half of the proteins encoded by newly sequenced genes (*2*, *3*). To improve this situation, new efforts in functional genomics and proteomics have been initiated. Among the proteomics efforts are ones in structural genomics. It is the objective of this paper to discuss new ways that nuclear magnetic resonance (NMR) can contribute to the structural genomics effort. NMR has to date contributed just 17% of the structures currently deposited in the Protein Data Bank (PDB),[1] compared to 82% by X-ray crystallography (*4*), but broad applicability to proteins which fail to provide diffraction quality crystals promises to increase the importance of contributions by NMR in the future.

Structural Genomics

Structural genomics, more properly structural proteomics, refers to the attempt to provide three-dimensional structural information about a significant fraction of the proteins encoded by the genes sequenced in various genome projects. The attempt is driven in part by a belief that three-dimensional structure will provide a better basis for recognition of function than sequence similarity (*5*−*8*). This belief is, in fact, supported by numerous examples of proteins of similar function that have little sequence homology but recognizable geometric similarities once functionally important residues are attached to specific points in a backbone fold (*9*, *10*). Given the possible utility of folds in these applications, as well as their use as a basis for homology modeling, most structural genomics efforts will not target all proteins in a genome, but will attempt to produce representative structures in each protein fold family. With fold families estimated to number from several thousand to ten thousand, this is still a daunting task, but one that may allow emerging computational methods to produce structures on a more global basis (*11*).

Meeting the challenge of structural genomics is now a worldwide effort with a large program underway in Japan (*12*) and new programs starting in Europe (*13*). Recently, the National Institute of General Medical Sciences (NIGMS), a part of the National Institutes of Health, announced the establishment of seven pilot centers in the United States for testing and establishing protocols for the large-scale production of protein structures (*14*). These new centers will build on a number of preexisting structural genomics projects in North America (*15*) and mesh with efforts to standardize

† This work was supported by grants from the National Science Foundation (MCB-9707728) and the National Institutes of Health (GM-62407).

[1] Abbreviations: PDB, Protein Data Bank; HSQC, heteronuclear single-quantum coherence; CLEANEX, clean chemical exchange spectroscopy; rmsd, root-mean-square deviation; BFP, blue fluorescent protein; 1D, 2D, and 3D, one-, two-, and three-dimensional, respectively; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; TROSY, transverse relaxation-optimized spectroscopy; COSY, coupling-correlated spectroscopy; DMPC, dimyristoylphosphatidylcholine; DHPC, dihexanoylphosphatidylcholine; DMPG, dimyristoylphosphatidylglycerol.

10.1021/bi0102095 CCC: $20.00
Published on Web 07/04/2001

procedures for capturing and storing both X-ray and NMR data (*16*−*18*).

Most of the new NIH centers will rely heavily on X-ray crystallography to produce structures. This is not surprising given the large percentage of the structures currently in the PDB that have been produced by X-ray crystallography. The advances in data collection efficiency at X-ray producing synchrotron sources will also ensure that X-ray crystallography will continue to be the major contributor to this database; however, one of the new centers does have NMR as a major component, and NMR is well represented in several others. NMR is also a major component in the Japanese and European efforts. As we discuss below, there is reason to believe that investment in this second contributor to protein structure determination will be productive.

NMR in Structural Genomics

There are several reasons to suggest that NMR will play a role in structural genomics activities. NMR can provide an important technique for selecting well-behaved proteins and optimizing conditions for structure determination, whether it be by NMR or X-ray crystallography. NMR can provide a means of identifying small ligands as well as macromolecular associations that may be essential for proper folding and function. And, NMR can provide an inherently complementary structure determination methodology. With respect to its complementarity, NMR structures are produced in solution, and there are a few documented cases where structural differences exist between solution and crystal forms of proteins (*19*−*21*). However, these cases are small in number, and differences are not likely to be significant at the level of structures for a protein fold database. A more important aspect is the possibility that NMR can provide structures for proteins that may be difficult to crystallize. For example, producing crystals of membrane proteins is still a challenge. Integral membrane proteins of the helical bundle class are estimated to represent 20−25% of most genomes (*22*). Yet, only ~1% of the structures deposited in the current PDB are classified as membrane proteins (*4*). There are NMR structures of several small proteins with transmembrane helical segments, some produced in micelle media by high-resolution techniques and some produced in more extended membrane mimetics using solids NMR techniques (*23*, *24*). Also, there are new techniques on the horizon that promise to make membrane proteins in micelles more accessible to high-resolution NMR (*25*). Several of the structural genomics pilot centers plan to initially exclude membrane proteins from consideration because of the difficulties in structure production by either NMR or X-ray crystallography. But membrane proteins will eventually have to be tackled, and they may not be the only class of proteins that prove to be difficult to crystallize.

It is widely recognized that heterogeneity in protein preparations (due to aggregation, for example) is detrimental to crystallization, and screening using light scattering is often used to avoid conditions that contribute to heterogeneous aggregation; both NMR and X-ray approaches suffer limitations due to aggregation. However, other factors are more detrimental to crystal-based approaches than NMR-based approaches to structure. For example, posttranslational modifications such as glycosylation seem to limit crystallizability, and disordered regions that may be integral parts of the protein structure seem to limit crystallizability (*26*). These latter factors do not normally prevent NMR-based structure determination of at least the well-structured regions.

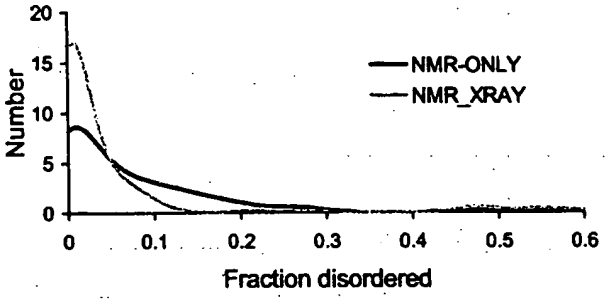


FIGURE 1: Influence of unstructured regions in proteins on the choice of structure determination approach. Data were taken from the July 1999 version of the Protein Data Bank.

It is possible to support an argument for the applicability of NMR in cases where crystallization has proved to be difficult by an examination of current versions of the PDB (*4*). There are now (January 2001) approximately 14 000 structures in the PDB; approximately 2200 have been determined by NMR methods. Most of these are deposited, not as a single structure, but as a set of 20 or more structures determined using random starting conditions in the structure determination protocol. Regions along the polypeptide backbone that are poorly defined can be located by calculating a root-mean-square deviation (rmsd) from the mean of the positions of backbone atoms residue by residue. We take an arbitrary limit of 4.0 Å as an indicator of poor structural definition. Poor definition can come simply from the absence of adequate NMR constraints, but the absence of constraints frequently correlates with motion or disorder in a region of the protein. Making this correlation, we can use large rmsds as an indicator of disordered regions in a protein.

Entries based on NMR (1504 from the July 1999 PDB) were filtered to produce a nonredundant set of proteins having more than 50 residues and adequate numbers of deposited structures (918). These entries were then evaluated on the basis of the fraction of residues showing an rmsd of >4.0 Å. The structures represented by these entries were divided into two sets on the basis of whether a corresponding X-ray structure existed. A search of all X-ray structures was conducted using a sequence identity search algorithm that allowed modest levels of amino acid substitution, deletion, or insertions (fewer than 12 residues different in size and fewer than 4 unmatched residues). On this basis, 182 were found to have a corresponding structure determined by X-ray crystallography. A search of the remaining NMR structures was conducted to select a set of comparable size having the least similarity to existing X-ray structures. A 163-member set having a BLAST (*27*) alignment homology score of less than 43 when compared to any X-ray structure was selected. The two subsets, NMR_ONLY and NMR_XRAY, were examined separately for the extent of disorder as given by the rmsd criterion applied to the NMR structure.

The analysis is presented graphically in Figure 1. For proteins with a low percentage of disorder, related X-ray structures are abundant, but for proteins with a higher percentage of disorder (more than 10%), the number of related X-ray structures drops off dramatically. NMR-based structures persist to a much higher percentage of disorder.

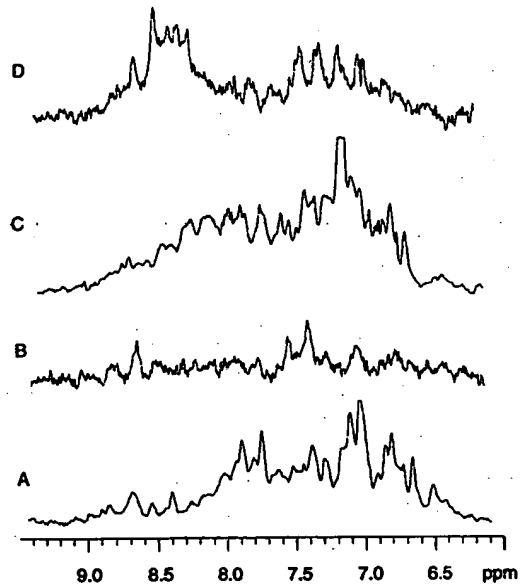


FIGURE 2: Screening for structured proteins using 1D amide proton exchange spectra. The sample is blue fluorescent protein at 0.5 mM: (A) 1D spectrum with Ca^{2+}, (B) CLEANEX spectrum with Ca^{2+}, (C) 1D spectrum without Ca^{2+}, and (D) CLEANEX spectrum without Ca^{2+}.

Such observations clearly support the suggestion that NMR-based structure determination is more applicable to proteins with disordered regions, possibly because these proteins may be difficult to crystallize.

NMR as a Screening Tool

Whether NMR is able to produce structures for disordered proteins, a simple ability to identify these proteins, or proteins that are heterogeneous because of aggregation or other conformational effects, would be a valuable contribution. Screening of expressed and purified proteins for sample conditions that are apt to promote crystallization or give good NMR samples is therefore a major activity of many of the pilot centers. We present in Figure 2 an illustration of the type of experiment that might be carried out on a very small amount of protein in a highly automated way. The experiment incorporates a simple test for rapidly exchanging amide protons. At pH 7, amide protons of unstructured regions of polypeptide chains undergo exchange with protons of water on time scales of tenths of seconds or less (*28*). In structured regions, amide protons are either buried in the hydrophobic interior of the protein or involved in hydrogen bonding. These amides exchange much more slowly (minutes to hours). A simple magnetization transfer experiment that uses magnetization associated with protons of water to provide detectable protein signal when they exchange into rapidly exchanging sites will selectively show amides in disordered regions. [In actual fact, a more sophisticated experiment that eliminates artifacts due to transfers from α-protons underlying the water resonance, CLEANEX, is used (*29*).]

In Figure 2, we show the amide proton regions of one-dimensional proton NMR spectra of a 20 kDa protein, obelin. Obelin is a photoprotein that emits light on addition of Ca^{2+} in a manner similar to that of its close relative, aquorin. The crystal structure of obelin has recently been determined, but only in the pre-emission form (*30*); it has not so far been possible to produce diffraction quality crystals of obelin in its reacted form, a form also known as blue fluorescent protein (BFP). Figure 2A shows the amide region of a simple 1D spectrum of BFP in the presence of Ca^{2+}; signals from all amides as well as a few aromatic protons are present. In Figure 2B, we show a spectrum of the same region produced using the CLEANEX experiment. As only signals that derive magnetization from protons on water can be detected in this experiment, those seen are from amide protons in moderately rapid exchange (exchange times of fewer than a few seconds) with water. The intensity in the 7.0−7.8 ppm region is seen for most proteins and includes intensity from side chain amide protons such as those on glutamine and asparagine. The few signals in the 8.1−8.9 ppm region are typical of a well-folded protein with few unstructured or surface-exposed amides. The spectrum suggests that at least in the presence of Ca^{2+} production of quality crystals may be possible. Spectra C and D of Figure 2 show a similar pair, but from the protein with Ca^{2+} removed. The intensity in the 8.1−8.9 ppm region is now abnormally high and indicative of partial unfolding of the backbone. The data suggest that the protein would be difficult to crystallize in the absence of Ca^{2+}, but the protein may be sufficiently folded for an NMR study. Screening based on experiments such as the CLEANEX experiment can be made quite efficient, and even automated, using NMR flow probes and micro manipulator robots. The spectra that are shown were acquired in approximately 15 min each on 0.5 mM protein samples using a standard 600 MHz spectrometer.

There is of course more to be gained from screens based on NMR spectra. Chemical shifts contain information. The appearance of the transfer intensity near 8 ppm is an example. This is a region characteristic of amides in random coil configuration. α-Proton chemical shifts in the 4−5 ppm region can provide equally valuable indicators of α-helix or β-strand structures, as can chemical shifts of the backbone α-carbons (*31*). Screening based on various types of NMR experiments has been adopted by several structural genomics projects (*32*, *33*). Screening is often based on 2D $^{15}N-^{1}H$ NMR, rather than the simple one-dimensional proton experiments described above. Resolution is greatly improved in 2D experiments but requires ^{15}N-labeled protein (this can be accomplished by growing *Escherichia coli* in minimal medium supplemented with [^{15}N]ammonium chloride at modest additional cost). The experiment employed is a workhorse 2D experiment called a heteronuclear single-quantum coherence (HSQC) experiment. The resulting spectrum shows cross-peaks, approximately one for each residue, that correlate chemical shifts of an amide proton with that of the directly bonded nitrogen. The experiment is efficient, requiring on the order of 15 min on a 1 mM protein sample using a modern 600 MHz NMR spectrometer. Positions of peaks can suggest disordered regions, and once the peaks are assigned to sequential positions, the information could actually be used to engineer proteins with the unfavorable segments removed. Also, given the normal count of approximately one cross-peak per residue, the detection of a surplus of peaks can indicate conformational heterogeneity that may adversely affect structure determination. The efficiency of the experiment also allows its use along with variation of pH or ionic strength or addition of cofactors in identification of conditions that produce minimal heterogene-

ity as well as optimum resolution for subsequent NMR experiments.

3D Structures by NMR

With the seeming advantages of NMR, why stop at screening; why not routinely continue on to complete structures? Obviously, there must be reasons if only 15−20% of the deposited structures in the PDB are NMR-derived. One reason that NMR has not played a larger role in structure determination is the limitation on sizes of proteins to which NMR methods can be applied. The limit on size stems from the need to resolve and assign resonances to particular proton sites in the protein sequence. Cross-peaks between assigned resonances in nuclear Overhauser effect spectroscopy (NOESY) ultimately give rise to the pairwise distance constraints used in most NMR structure determinations (*34*). Resolution depends inversely on resonance line widths, and the widths depend in turn on effective molecular weight. Limits imposed by resolution have been pushed back over the years, first by use of isotope enrichment (^{13}C, ^{15}N, and ^{2}H) (*35*) and extension of spectra to three and four dimensions and more recently by transverse relaxation optimization techniques (TROSY) in combination with high magnetic fields (*25*). However, complete structure determination of monomeric proteins still has not pushed beyond a molecular mass limit of 50 kDa, and determinations not requiring deuteration of the protein still seem to be limited to 25−30 kDa (*33*). While this is a severe limitation, it is mitigated by the fact that the average domain size of encoded proteins appears to be only slightly greater than 150 amino acids, or 17 kDa (*3*).

A more severe limitation is that the time required for data acquisition and analysis is long, and sample preparation requires the use of isotopically labeled media (^{15}N- and ^{13}C-labeled proteins). There have been enormous strides made in the efficient production of proteins through expression in *E. coli* (*32*), and new cell-free production techniques pioneered in Japan promise more latitude in produced proteins and incorporated labels (*12*). However, the 4−6 weeks of acquisition and subsequent months-long periods required for assignment and structure determination is still a major obstacle (*33*). This time scale is not compatible with structural genomics objectives that would require 100−200 structures per year from each of the seven NIH-sponsored pilot centers (*14*).

Automation of NOE-Based Methods

One approach to reducing the time requirements relies on extensive automation to reduce the analysis time and improved instrument hardware to reduce data acquisition time. Isotopic labeling of proteins with both ^{13}C and ^{15}N has led to a very reliable and automatable assignment strategy. Primary experiments use one bond scalar coupling connectivities to walk along the peptide backbone. Extending connectivities out to the β carbons of the amino acid side chains allows assignment of many sequentially connected residues to amino acid types based on chemical shift correlations. Inclusion of experiments such as HCCH TOCSY that use isotropic mixing sequences to make multiple bond connectivities through the amino acid side chains allows assignment of most carbon and proton resonances. The program AUTOASSIGN developed by the Rutgers group is representative of programs that provide a very efficient automation of much of this general strategy (*36*). While AUTOASSIGN and most other automated assignment programs rely on ^{13}C and ^{15}N isotopic labeling (*37*), some can be used to aid assignment even in the absence of ^{13}C labeling (*38*−*40*). These alternate strategies may become important for proteins that are difficult to express in bacterial systems or cell-free systems.

Assignment of resonances, however, still falls short of assignment of peaks in NOESY spectra. This is a particularly challenging task because of the severe overlap and resulting ambiguities in assignment. Automated methods for this step fortunately also exist (*41*, *42*). They mesh with commonly used simulated annealing protocols for structure determination and allow iterative determinations for elimination of ambiguities in assignment. All of these automated assignment tools are still very much in the development stage. The number of structures produced with the aid of automation is at this point small, but structural genomics efforts will certainly require implementation of these techniques. In a recent example that used some of the automated tools that are mentioned, time from receipt of a plasmid containing a 90-residue soluble (2−3 mM) protein to production of a tertiary fold was reduced to <1 month (*43*).

Very recently, the promise of reducing actual data collection time has come from the introduction of NMR cryoprobes in which the receiver coil and associated electronic components are cooled to very low temperatures (*33*). Sensitivity improvements on protein samples of a factor of ~3 can be achieved. This translates to a savings in time of a factor of 9 when substantial signal averaging is required. A recent example on a protein of 180 residues shows a reduction to 1.5 days for experiments required for backbone assignments (*44*).

Substantial savings can also be achieved in acquisition and analysis of experiments needed for side chain assignments and NOE cross-peak analysis by selective protonation of groups in otherwise fully deuterated proteins. The object is to reduce the NOE peaks to a set very rich in useful distance constraints, often a set containing peaks from side chains of residues concentrated at the hydrophobic core of proteins. This strategy was first well-illustrated in the work of Gardner and Kay where methyl groups were selectively protonated (*45*). The strategy has recently been extended by Fesik and co-workers and combined with the use of a cryoprobe to acquire sufficient data for the determination of the fold of a 180-residue protein in just 4 days (*44*). The procedure does, however, require preparation of several samples with different distributions of isotopic labels, and it does require the use of amino acids synthesized to have ^{13}C and/or protons in specific places.

Applications of NOE-Based Methods

It is useful to examine at least one recent example of using the triple-resonance assignment, NOE-based structure determination, approach in a structural genomics application. We choose an application to MHT538, a protein of approximately 110 amino acids from the thermophilic archaeon *Methanobacterium thermoautotrophicum* (*46*). Aside from efficiency issues, which are not discussed in detail, this

application includes a nice illustration of additional ways that NMR might contribute in pushing past structure, to the point of functional characterization. The protein target was chosen without hypothesis as to function in an attempt to see what could be learned from the structure itself. A PSI-BLAST search (*27*) using the MHT538 sequence suggested a weak similarity (18−23% for a 95-residue segment) to members of a family of proteins annotated as putative ATPases or kinases (COG1618). As this level of similarity would normally be inadequate for prediction of any structural homology, an NMR-based structure determination proceeded using ^{13}C- and ^{15}N-labeled protein and a battery of experiments similar to that described above. A structure falling into the fold family $(\alpha/\beta)^5$ (minus the helix between strands 2 and 3) was determined with a 0.88 Å rmsd for backbone atoms of residues 4−108. This structure does not, in fact, bear a relationship to structures found for other members of the sequence family of putative ATPases or kinases. Using tools for finding structural homologies [SCOP (*47*) and DALI (*48*)], a relationship to either flavodoxins, or a receiver domain of two-component response regulator systems such as that of CheY, was found.

At this point, NMR screening methodology was used to draw a distinction between these two functional classification choices. As described previously, the simple two-dimensional $^{15}N-^{1}H$ heteronuclear single-quantum coherence (HSQC) spectrum provides a map of the protein backbone with approximately one cross-peak for each amino acid (each H−N amide pair). Movement of specific cross-peaks on binding of ligands to proteins is now used extensively in the pharmaceutical industry to identify binding sites on the protein surface (*49*). In the MTH538 study, this technique was used to explore possible flavin interactions, as expected if the protein were a flavodoxin, and possible Mg^{2+} binding, as expected for receiver domains; no flavin binding was detected, but Mg^{2+} binding to a site homologous to one in CheY was found. Although some other characteristics of receiver domains were missing, the close structural homology, along with the specific Mg^{2+} site, was used to suggest a possible receiver domain function.

Fold Determination Methods

Despite the advances in the efficiency of traditional NMR approaches to structure determination, and the successful illustrations of the contributions that these NMR approaches can make to structural genomics, the rate of data production still pales in comparison to that of synchrotron-based X-ray crystallography. This observation suggests that some reconsideration of both the objectives and approaches used in NMR applications to structural genomics projects might prove to be useful. First, in terms of an objective, is the traditional complete high-resolution structure appropriate for structural genomics applications? High-resolution structures will always be an important part of detailed mechanistic investigations, but functional classification of proteins may be a different matter. One of the principal hopes of structural genomics is that representative structures in each of a few thousand fold families will provide a basis upon which computational biologists can produce structures of related proteins and draw conclusions about function. When homology modeling is used on systems that have sequences that are as little as 25% identical, one clearly does not use the precise placement of either backbone or side chain atoms of representative structures in a fold library. Also, functional identification algorithms have been devised that work on the basis of approximate placement of backbone atoms as opposed to precise placement of side chain atoms (*6*). In the future, new computational methods for placing side chains into backbone structures promise to make backbone structures even more useful (*50*, *51*). In these cases, placement of backbone atoms with accuracy that is better than that which can be achieved by homology modeling appears to be necessary, and new experimental methods directed at accurate backbone structure determination will be useful.

If backbone structures become a primary objective, the traditional NOE-based NMR approach to structure can also be questioned. This approach is not optimal for direct investigation of backbone structures because the short-range character of the NOE, from which distance constraints are derived, dictates that side chain−side chain contacts in the hydrophobic core play an important part in determining a protein fold. However, a source of structural data that can constrain more directly backbone atom positions has come on the scene recently, residual dipolar couplings (*52*, *53*). The same dipolar interaction that gives rise to the NOE actually contains both angle- and distance-dependent terms. In the NOE, the angle-dependent part is responsible for modulation of the interaction as a protein tumbles in solution and is essential for making the distance dependence of NOEs measurable as a spin relaxation phenomenon. However, it normally makes no direct contribution to spin state energy differences that might be efficiently measured through variations in resonance positions. This is because the angular term rigorously averages to zero over the time course of molecular tumbling in isotropic solutions. Measurements of residual dipolar couplings depend on restoring small levels of directional order to proteins as they tumble in solution (of order one part in a thousand). This was first done for NMR observations on proteins using the inherent tendency of paramagnetic proteins to orient in high magnetic fields (*54*), but since that time, the use of aqueous liquid crystal media such as bicelles (*53*), phage (*55*, *56*), or cellulose microcrystals (*57*) has become standard practice.

The introduction of order leads to a residual contribution to the splitting of resonance multiplets, D^{res}_{ij}. In the case of a directly bonded pair of spin $^1/_2$ nuclei ($^1H-^{15}N$, $^1H-^{13}C$, etc.), where the internuclear distance is known, the contribution to splitting becomes a simple function of order and orientation adopted by the system. The effect of order is in turn conveniently expressed in terms of elements of an order tensor (S_{ij}) and direction cosines relating the internuclear vector to a principal order frame [$\cos(\alpha_i)$] (*58*):

$$D^{res}_{ij} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_i\gamma_j h}{2\pi^2 r_{ij}^3}\sum_{kl} S_{kl}\cos(\alpha_k)\cos(\alpha_l) \qquad (1)$$

Despite the apparent complexity of the equation, it has only five independent parameters; these can alternately be described in terms of three Euler angles relating the orientation of a molecular fragment to a principle alignment frame and principal and rhombic alignment parameters (*53*). The latter two apply to all parts of a rigid protein molecule. The first

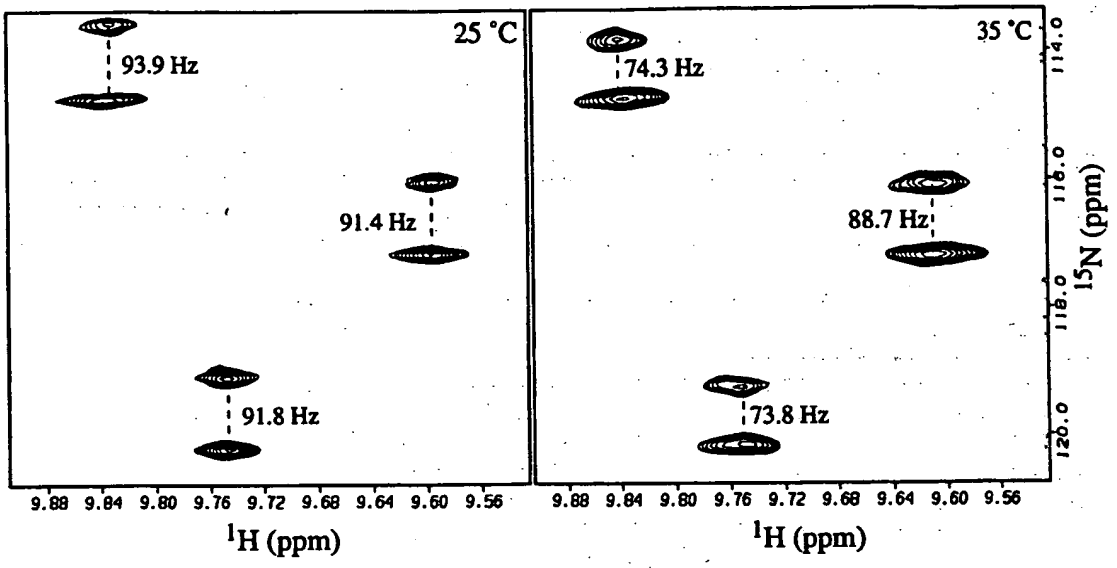


FIGURE 3: Measurement of ^{15}N−^{1}H residual dipolar couplings in amide groups using coupled HSQC spectra. The sample is 0.3 mM ADP ribosylating factor (ARF) in a 7% bicelle preparation (2.8:1 DMPC/DHPC or 15:1 DMPC/DMPG) at pH 6.5 (100 mM NaCl, 1 mM $MgCl_2$, and 10 mM NaH_2PO_4). The sample is isotropic at 25 °C and ordered at 35 °C.

three can be determined independently for any semirigid fragment in a protein, and can serve to relate orientations of identifiable structural elements.

Part of the appeal of residual dipolar couplings measurements for structural genomics applications is the efficiency with which these data can be acquired. Measurement of ^{15}N−^{1}H dipolar couplings of directly bonded amide pairs is based on the very efficient HSQC experiment described above. Modifications necessary for the measurement of residual dipolar couplings can be as simple as removing the normal 180 pulse used for refocusing proton scalar coupling during the indirect detection period. When this is done, the single cross-peak for each amide turns into a doublet with a splitting at the sum of one-bond scalar and dipolar contributions. The dipolar part is usually extracted by acquiring spectra under both isotropic and oriented conditions. As illustrated in Figure 3 for an ^{15}N-labeled sample of a 20 kDa protein, ADP ribosylating factor, this can be simply done using bicelle systems in which isotropic and ordered states can be produced by changing the temperature from 25 to 35 °C. At 25 °C, only scalar couplings contribute, and splittings are all of a similar magnitude. At 35 °C, the bicelles, which are simply discoidal pieces of lipid bilayer, order with normals perpendicular to the magnetic field. Occasional collisions of proteins with the ordered surfaces impart a small degree of order; the dipolar contributions to couplings between ^{1}H and ^{15}N nuclei fail to average to zero, and splittings increase or decrease depending on the angles that a particular internuclear vector make with the axes of the order frame. There are of course a number of more sophisticated pulse sequences that allow measurement of ^{1}H−^{15}N and ^{1}H−^{13}C couplings from either intensity variation or frequency variation of cross-peaks (*59*, *60*). A primary advantage of the intensity-based sequences is that they encode couplings in single cross-peaks that can have the same positions as the cross-peaks in standard HSQC and triple-resonance experiments. There have also been 3D experiments devised that allow measurement of several couplings at once (*61*). For backbone couplings (^{1}H−^{15}N amide, ^{1}H−$^{13}C_\alpha$, ^{1}HN−^{13}CO, $^{13}C_\alpha$−^{15}N, ^{13}CO−^{15}N, etc.), it is important to realize that these involve atoms that are among the most easily assigned using standard triple-resonance experiments and automated assignment programs. Moreover, orientational relationships of fragments containing these atoms can be defined without close approach of the fragments. Hence, definition of a backbone structure without simultaneous placement of side chains is possible.

There has been a wide range of applications of residual dipolar couplings to protein structure determination, including their use to refine structures that have been produced using substantial amounts of data from NOE distance restraints (*62*). However, the appeal in structural genomics results from the possibility of more direct use in structure recognition or structure determination. The fact that ^{1}H−^{15}N dipolar couplings could be used in properly classifying a structure using a fold library was clearly illustrated in work by Annila et al. (*63*). Here a protein called calerythrin (~180 residues) was labeled with ^{15}N and ^{13}C to assign backbone resonances, and ^{15}N−^{1}H dipolar couplings for amide bonds were measured. Experimental couplings were then compared to couplings predicted for a small set of known structures. The measured sequence was threaded into each structure using secondary structure information from backbone carbon chemical shifts to improve threading, and coupling sets were generated for a grid of possible order frame alignments. The set of structures included two sarcoplasmic calcium binding proteins, one from sandworm and one from an amphioxus, whose sequences were 27 and 15% identical, respectively. These proved to be the only structures giving a good match to measured dipolar coupling patterns. Hence, a strong argument for an ability to place new proteins into existing fold classes strictly on the basis of easily acquired residual dipolar coupling data resulted. Procedures for searching databases for homologous folds based on dipolar data, secondary structure information, and pseudocontact shifts accessible for paramagnetic proteins have been efficiently

programmed by other groups. The Griesinger group reports success with three test proteins using a more extended, but still small, fold database having 125 members (*64*).

A variation of the above procedure that can extend structure prediction to proteins representing new folds has also been described (*65*). The method, termed molecular fragment replacement, does not require that an entire fold be represented in a database, but only short segments of structure (approximately seven residues in the application described). Residual dipolar data on the small protein ubiquitin (76 residues) that included $^1H-^{15}N$ amide, $^1H-^{13}C_\alpha$,$^{13}CO-^{15}N$[1], and $^1HN-^{13}CO$ couplings obtained in two different aligned media were used. Seven-residue segments were threaded trough structures in a reduced PDB (1560 proteins), back calculating couplings from optimized alignment parameters at each step. Matching couplings produced local structures that could be assembled, and the resulting complete structure could be refined against the residual dipolar and chemical shift data. A backbone model, excluding a flexible C-terminus, matches the crystal structure of ubiquitin to within 0.88 Å. Again, this procedure employs only data that can be easily and efficiently acquired.

Direct calculation of backbone structures, without the use of a database, and using primarily residual dipolar data can also be accomplished. In our own work, two small proteins were used as test cases, acyl carrier protein (ACP) from *E. coli*, a protein of 77 amino acids whose structure had previously been characterized by traditional NMR methods, and NodF, a 92-amino acid product of a *Rhizobium leguminosarum* gene required for the synthesis of molecules that stimulate symbiotic root nodule formation (*66*). The acyl carrier protein test was carried out without the benefit of ^{13}C isotopic labeling (only ^{15}N labeling was used). Because ^{15}N-enriched media can be prepared at a fraction of the cost of preparing ^{15}N- and ^{13}C-enriched media, this may be important for gene products that give poor yields when expressed using *E. coli*, or require expression using other organisms. Elements of secondary structure were identified from a combination of backbone NOE patterns and three-bond ($^3J_{HN-HA}$) scalar couplings. These elements, three α-helices, were modeled as polyalanine helices of ideal geometry, and the orientation of each helix relative to a global orienting frame was determined by inverting equations similar to eq 1. The residual dipolar data came from $^1H-^{15}N$ couplings measured in HSQC-based experiments and $^1H-^1H$ couplings measured in a ^{15}N-edited constant time COSY experiment. A very small number of easily assigned NOEs connecting to backbone nuclei (five) were used to help restrict translational degrees of freedom for the helices. The resulting structure agreed with the previously determined NMR structure to within 3 Å for backbone atoms of the secondary structure elements (see Figure 4). This level of resolution is not representative of the precision of the measurements but is limited by the attempt to fit real helices with idealized models. Allowing helices to bend greatly improves the fit to data and presumably the accuracy of the structures (*66*). More importantly for structural genomics applications, the total data acquisition time was estimated to be <1 week on a conventional 500 MHz spectrometer. With the benefit of cryoprobes and high-field spectrometers, this should be reduced to approximately 1 day.

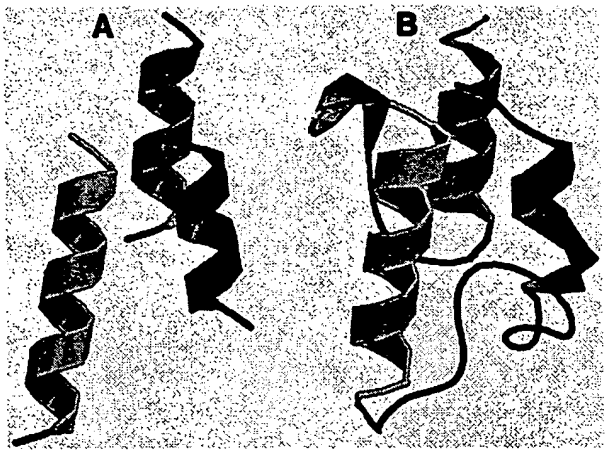


FIGURE 4: Structure of the acyl carrier protein (ACP) determined from residual dipolar couplings (A) and from NOEs (B).

The NodF protein was doubly labeled and, hence, benefited from a more efficient assignment strategy and the measurement of additional residual dipolar couplings ($^1H-^{13}C$, $^{13}C'-^1HN$, and $^{13}C'-^{15}N$ couplings, in addition to the couplings used in the ACP case). The additional couplings allowed modeling with helices assembled from several canonical segments (to allow bending). Data acquisition was actually shorter than for the ^{15}N-labeled ACP sample, approximately 5 days. No other structure is available for this protein, so no evaluation of the accuracy of the structure could be made.

Other procedures for more directly building backbone structures of protein from residual dipolar data are being developed by other groups (*67*, *68*). Some of these strategies are based on an ability to treat peptide planes as independent fragments and orient them one plane at a time. They use an extensive set of residual dipolar data from doubly labeled proteins, but no or little NOE data. Data acquisition time is again short, and will be further reduced with the aid of cryogenic probes.

Future of NMR in Structural Genomics

With advances such as the ones described above, we can remain optimistic about the role NMR will play in structural genomics. It can provide structures for proteins that may not easily crystallize, and it can provide these structures rapidly, especially if specification of backbone atoms in a protein fold is the primary objective. It is also useful to look beyond completion of an appropriate fold database to functional characterization of proteins. NMR will play a role here as well. We have already mentioned the potential of HSQC spectra in screens for ligand binding and drug design (*49*). It is important to realize that these experiments can also be used in screens for protein−protein interactions (*69*, *70*). It is now clear that relatively few proteins act in isolation, and understanding function may require looking at combinations of proteins. Going beyond simple detection of interaction, perturbations of specific peaks in HSQC experiments can identify regions of contact. The relative orientation of proteins in contact can also be determined using the residual dipolar methods mentioned above. Entire proteins of known structure can be treated as orientable fragments, just as we did for secondary structure elements or individual peptide

planes in the work described above. In fact, there are now several examples of the determination of the relative orientation of loosely connected domains in multiple domain proteins using residual dipole methods (*71*−*73*). In one of these examples, binding of ligands between domains makes the relative orientations in the presence and absence of ligand an integral part of understanding the mechanism (*72*). In a more hypothetical case, cooperative binding of carbohydrate ligands on the surfaces of membranes may be modulated by the relative orientation of subunits in multimeric lectins (*74*). In short, prospects for contributions beyond the initial focus of structural genomics initiatives on domain structures are bright. We look forward to an ability to report progress in this direction in a few years.

ACKNOWLEDGMENT

We thank Dr. C. A. Fowler for his analysis of the ACP structure and Drs. J. W. Lee and E. Vysotski for their collaboration on the obelin samples.

REFERENCES

1. Sali, A. (1998) *Nat. Struct. Biol. 5*, 1029−1032.
2. Blundell, T. L., and Mizuguchi, K. (2000) *Prog. Biophys. Mol. Biol. 73*, 289−295.
3. Burley, S. K. (2000) *Nat. Struct. Biol. 7*, 932−934.
4. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res. 28*, 235−242.
5. Orengo, C. A., Todd, A. E., and Thornton, J. M. (1999) *Curr. Opin. Struct. Biol. 9*, 374−382.
6. Skolnick, J., and Fetrow, J. S. (2000) *Trends Biotechnol. 18*, 34−39.
7. Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000) *Nat. Struct. Biol. 7*, 986−990.
8. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000) *Nat. Struct. Biol. 7*, 991−994.
9. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M., and Arrowsmith, C. H. (2000) *Nat. Struct. Biol. 7*, 903−909.
10. Skolnick, J., Fetrow, J. S., and Kolinski, A. (2000) *Nat. Biotechnol. 18*, 283−287.
11. Gerstein, M. (2000) *Nat. Struct. Biol. 7*, 960−963.
12. Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., and Kuramitsu, S. (2000) *Prog. Biophys. Mol. Biol. 73*, 363−376.
13. Heinemann, U. (2000) *Nat. Struct. Biol. 7*, 940−942.
14. Norvell, J. C., and Machalek, A. Z. (2000) *Nat. Struct. Biol. 7*, 931.
15. Terwilliger, T. C. (2000) *Nat. Struct. Biol. 7*, 935−939.
16. Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000) *Nat. Struct. Biol. 7*, 957−959.
17. Brunger, A. T., and Laue, E. D. (2000) *Curr. Opin. Struct. Biol. 10*, 557.
18. Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E., and Wuthrich, K. (1998) *Eur. J. Biochem. 256*, 1−15.
19. Wuthrich, K. (1995) *Acta Crystallogr. D51*, 249−270.
20. Clore, G. M., Bax, A., Ikura, M., and Gronenborn, A. M. (1993) *Curr. Opin. Struct. Biol. 3*, 838−845.
21. Engh, R. A., Dieckmann, T., Bode, W., Auerswald, E. A., Turk, V., Huber, R., and Oschkinat, H. (1993) *J. Mol. Biol. 234*, 1060−1069.
22. von Heijne, G. (1999) *Q. Rev. Biophys. 32*, 285−307.
23. Sanders, C. R., and Oxenoid, K. (2000) *Biochim. Biophys. Acta 1508*, 129−145.
24. de Groot, H. J. M. (2000) *Curr. Opin. Struct. Biol. 10*, 593−600.
25. Riek, R., Pervushin, K., and Wuthrich, K. (2000) *Trends Biochem. Sci. 25*, 462−468.
26. Kwong, P. D., Wyatt, R., Desjardins, E., Robinson, J., Culp, J. S., Hellmig, B. D., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1999) *J. Biol. Chem. 274*, 4115−4123.
27. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res. 25*, 3389−3402.
28. Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997) *Protein Sci. 6*, 1101−1109.
29. Hwang, T. L., Mori, S., Shaka, A. J., and vanZijl, P. C. M. (1997) *J. Am. Chem. Soc. 119*, 6203−6204.
30. Vysotski, E. S., Liu, Z. J., Rose, J., Wang, B. C., and Lee, J. (1999) *Acta Crystallogr. D55*, 1965−1966.
31. Wishart, D. S., and Sykes, B. D. (1994) *Methods Enzymol. 239*, 363−392.
32. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C. H., and Edwards, A. M. (2000) *Prog. Biophys. Mol. Biol. 73*, 339−345.
33. Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., and Szyperski, T. (2000) *Nat. Struct. Biol. 7*, 982−985.
34. Wider, G., and Wuthrich, K. (1999) *Curr. Opin. Struct. Biol. 9*, 594−601.
35. Gardner, K. H., and Kay, L. E. (1998) *Annu. Rev. Biophys. Biomol. Struct. 27*, 357−406.
36. Zimmerman, D. E., Kulikowski, C. A., Huang, Y. P., Feng, W. Q., Tashiro, M., Shimotakahara, S., Chien, C. Y., Powers, R., and Montelione, G. T. (1997) *J. Mol. Biol. 269*, 592−610.
37. Moseley, H. N. B., and Montelione, G. T. (1999) *Curr. Opin. Struct. Biol. 9*, 635−642.
38. Oschkinat, H., and Croft, D. (1994) *Methods Enzymol. 239*, 308−318.
39. Bartels, C., Guntert, P., Billeter, M., and Wuthrich, K. (1997) *J. Comput. Chem. 18*, 139−149.
40. Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K. P., and Kalbitzer, H. R. (2000) *J. Biomol. NMR 17*, 137−151.
41. Xu, Y., Wu, J., Gorenstein, D., and Braun, W. (1999) *J. Magn. Reson. 136*, 76−85.
42. Nilges, M., Macias, M. J., Odonoghue, S. I., and Oschkinat, H. (1997) *J. Mol. Biol. 269*, 408−422.
43. Kozlov, G., Ekiel, I., Beglova, N., Yee, A., Dharamsi, A., Engel, A., Siddiqui, N., Nong, A., and Gehring, K. (2000) *J. Biomol. NMR 17*, 187−194.
44. Medek, A., Olejniczak, E. T., Meadows, R. P., and Fesik, S. W. (2000) *J. Biomol. NMR 18*, 229−238.
45. Gardner, K. H., Rosen, M. K., and Kay, L. E. (1997) *Biochemistry 36*, 1389−1401.
46. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H., and Kennedy, M. A. (2000) *J. Mol. Biol. 302*, 189−203.
47. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol. 247*, 536−540.
48. Holm, L., and Sander, C. (1999) *Nucleic Acids Res. 27*, 244−247.
49. Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1999) *Q. Rev. Biophys. 32*, 211−240.
50. Moult, J., and Melamud, E. (2000) *Curr. Opin. Struct. Biol. 10*, 384−389.
51. Looger, L. L., and Hellinga, H. W. (2001) *J. Mol. Biol.* (submitted for publication).
52. Prestegard, J. H. (1998) *Nat. Struct. Biol. 5*, 517−522.
53. Tjandra, N., and Bax, A. (1997) *Science 278*, 1111−1114.
54. Tolman, J. R., Flanagan, J. M., Kennedy, M. A., and Prestegard, J. H. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 9279−9283.
55. Hansen, M. R., Mueller, L., and Pardi, A. (1998) *Nat. Struct. Biol. 5*, 1065−1074.
56. Clore, G. M., Starich, M. R., and Gronenborn, A. M. (1998) *J. Am. Chem. Soc. 120*, 10571−10572.

57. Fleming, K., Gray, D., Prasannan, S., and Matthews, S. (2000) *J. Am. Chem. Soc. 122*, 5224−5225.
58. Prestegard, J. H., Al-Hashimi, H. M., and Tolman, J. R. (2001) *Q. Rev. Biophys.* (in press).
59. Ottiger, M., and Bax, A. (1998) *J. Biomol. NMR 12*, 361−372.
60. Tolman, J. R., and Prestegard, J. H. (1996) *J. Magn. Reson., Ser. B 112*, 245−252.
61. Wang, Y. X., Marquardt, J. L., Wingfield, P., Stahl, S. J., Lee-Huang, S., Torchia, D., and Bax, A. (1998) *J. Am. Chem. Soc. 120*, 7385−7386.
62. Clore, G. M., Starich, M. R., Bewley, C. A., Cai, M. L., and Kuszewski, J. (1999) *J. Am. Chem. Soc. 121*, 6513−6514.
63. Annila, A., Aitio, H., Thulin, E., and Drakenberg, T. (1999) *J. Biomol. NMR 14*, 223−230.
64. Meiler, J., Peti, W., and Griesinger, C. (2000) *J. Biomol. NMR 17*, 283−294.
65. Delaglio, F., Kontaxis, G., and Bax, A. (2000) *J. Am. Chem. Soc. 122*, 2142−2143.
66. Fowler, C. A., Tian, F., Al-Hashimi, H. M., and Prestegard, J. H. (2000) *J. Mol. Biol. 304*, 447−460.
67. Mueller, G. A., Choy, W. Y., Yang, D. W., Forman-Kay, J. D., Venters, R. A., and Kay, L. E. (2000) *J. Mol. Biol. 300*, 197−212.
68. Hus, J. C., Marion, D., and Blackledge, M. (2001) *J. Am. Chem. Soc. 123*, 1541−1542.
69. Sette, M., Spurio, R., Van Tilborg, P., Gualerzi, C. O., and Boelens, R. (1999) *RNA 5*, 82−92.
70. Rajagopal, P., Waygood, E. B., Reizer, J., Saier, M. H., and Klevit, R. E. (1997) *Protein Sci. 6*, 2624−2627.
71. Bewley, C. A., and Clore, G. M. (2000) *J. Am. Chem. Soc. 122*, 6009−6016.
72. Skrynnikov, N. R., Goto, N. K., Yang, D. W., Choy, W. Y., Tolman, J. R., Mueller, G. A., and Kay, L. E. (2000) *J. Mol. Biol. 295*, 1265−1273.
73. Fischer, M. W. F., Losonczi, J. A., Weaver, J. L., and Prestegard, J. H. (1999) *Biochemistry 38*, 9013−9022.
74. Imberty, A., and Drickamer, K. (1999) *Curr. Opin. Struct. Biol. 9*, 547−548.

BI0102095

EXHIBIT 5

NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230

DIVISION OF MOLECULAR AND CELLULAR BIOSCIENCES

Dr. Thomas A. Szyperski
Department of Chemistry
SUNY at Buffalo - Amherst Campus
816 Natural Sciences Complex
Buffalo, NY 14260

MAY 5 2000

Ref: MCB-0075773

Dear Dr. Szyperski:

We regret to inform you that the National Science Foundation is unable to support your proposal entitled "Reduced Dimensionality NMR Spectroscopy for Structural Genomics".

A number of factors are considered in evaluating each proposal submitted to the Foundation. Of these, scientific merit is most important. Other factors include the relation of the proposed research to other research in the area and the distribution of funds among the various areas of biological sciences. Many meritorious proposals cannot be funded simply because of the limited resources available.

For further information concerning the evaluation of your proposal, please contact the Program Officer whose name and telephone number appears on the enclosed *Context Statement.* Copies of the reviews of your proposal and a *Panel Summary,* outlining the salient points raised in the panel discussion of your proposal, are enclosed. These are for your personal use and are not made available by the Foundation to anyone else. They may be helpful to you in preparing future proposals.

Although we were unable to support this proposal, we would be pleased to consider any future proposals you may wish to submit.

Sincerely,

Christopher Greer, Ph.D.
Acting, Deputy Division Director

Enclosures
Copy (without reviews) to:
Kurt Winter
Grant & Contract Admin.

Panel Summary
Molecular Biophysics
Szyperski, Thomas
MCB 0075773

The PI proposes to further develop methods to reduce the time for structure determination by NMR. This work is based on some nice experiments that the PI has already published. The PI projects that implementation of an RD NMR package would result in time savings for a facility trying to maximize throughput for protein NMR. Unfortunately, the project is essentially for the generation of software with no development of science. The NMR community has not appeared so far to be interested in RD. The panel concluded that new demonstrations of practicality of RD NMR were needed.

Panel rating: Good

EXHIBIT 6

Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment

Thomas Szyperski*†, Deok C. Yeh*‡, Dinesh K. Sukumaran*, Hunter N. B. Moseley§, and Gaetano T. Montelione§

*Departments of Chemistry and Structural Biology, State University of New York, Buffalo, NY 14260; and §Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854

Communicated by Herbert Hauptman, Hauptman–Woodward Medical Research Institute, Buffalo, NY, April 12, 2002 (received for review November 15, 2001)

A suite of reduced-dimensionality $^{13}C,^{15}N,^{1}H$-triple-resonance NMR experiments is presented for rapid and complete protein resonance assignment. Even when using short measurement times, these experiments allow one to retain the high spectral resolution required for efficient automated analysis. "Sampling limited" and "sensitivity limited" data collection regimes are defined, respectively, depending on whether the sampling of the indirect dimensions or the sensitivity of a multidimensional NMR experiments *per se* determines the minimally required measurement time. We show that reduced-dimensionality NMR spectroscopy is a powerful approach to avoid the "sampling limited regime"—i.e., a standard set of ten experiments proposed here allows one to effectively adapt minimal measurement times to sensitivity requirements. This is of particular interest in view of the greatly increased sensitivity of NMR spectrometers equipped with cryogenic probes. As a step toward fully automated analysis, the program AUTOASSIGN has been extended to provide sequential backbone and $^{13}C^{\beta}$ resonance assignments from these reduced-dimensionality NMR data.

Rapid resonance assignment is a prerequisite for high-throughput (HTP) structure determination and structural genomics (1). The aims of structural genomics are to (*i*) explore the naturally occurring "protein fold space" and (*ii*) contribute to the characterization of function through the assignment of atomic-resolution three-dimensional (3D) structures to proteins. The ultimate goal is to provide one or more representative 3D structures for every structural domain family in nature. It is now generally acknowledged that NMR will play an important role in this endeavor (1). The resulting demand for HTP structure determination requires fast and automated NMR data collection and analysis protocols. This impetus for the development of new methods will have broad impact in the technological infrastructure for structural biology and molecular biophysics.

Two key objectives for NMR data collection can be identified. Firstly, the measurement time should be minimized so as to lower the cost per structure and relax the constraint that NMR samples need to be stable over long time periods. Secondly, automated analysis requires recording of a redundant set of NMR spectra each affording good resolution, while it is also desirable to keep the total number of spectra small to reduce complications due to interspectral variations of chemical shifts (2). This second objective can be addressed by maximizing the dimensionality of the spectra. However, the joint realization of the first and second objective is impeded by the large lower bounds for measurement times of four (or higher) dimensional NMR spectra arising from the independent sampling of three (or more) indirect dimensions.

We distinguish "sampling limited" and "sensitivity limited" data collection regimes, depending on whether the sampling of the indirect dimensions or the sensitivity of the multidimensional NMR experiments *per se* determines the minimally achievable measurement time. Because structure determinations rely on nearly complete shift assignments routinely obtained using $^{13}C,^{15}N,^{1}H$-triple-resonance (TR) NMR (3), the development of techniques that avoid the sampling limited regime represents an important challenge. Reduced-dimensionality (RD) TR NMR experiments (4–7), designed for simultaneous frequency labeling of two spin types in a single indirect dimension, offer a viable strategy to circumvent sampling-limited recording of NMR spectra. RD NMR is based on a projection technique for reducing the spectral dimensionality: the chemical shifts of the projected dimension give rise to a cosine-modulation of the transfer amplitude, yielding peak doublets encoding n chemical shifts in an n-1 dimensional spectrum. Thus, for example, four-dimensional (4D) information can be obtained in a 3D experiment. This reduces the sampling requirements and the minimal measurement time by about an order of magnitude (7), which allows recording projected 4D experiments within a few hours while retaining maximal evolution times and thus a resolution routinely achieved in conventional 3D NMR spectra. Furthermore, axial coherences, arising from either incomplete polarization transfer or steady-state heteronuclear magnetization, can be observed as peaks located at the center of the doublets (6). This allows both the unambiguous assignment of multiple doublets with degenerate chemical shifts in the other dimensions and the identification of cross peak pairs by symmetrization of spectral strips about the position of the central peak. RD NMR experiments were the first designed to simultaneously recruit both ^{1}H and heteronuclear magnetization for signal detection (6), and RD two-spin coherence NMR spectroscopy (8) serves as a valuable radio-frequency (rf) pulse module for measurement of cross-correlated heteronuclear relaxation rates (9). Here we present a suite of nine RD TR NMR experiments (six of which are unique implementations) for complete protein resonance assignment. To integrate the RD NMR technology for rapid assignment, we have, as a step toward fully automated analysis, extended the program AUTOASSIGN (10, 11) for the use of RD NMR spectra.

Materials and Methods

NMR measurements were performed at 25°C on a Varian Inova 600 spectrometer by using a 1-mM solution of $^{13}C/^{15}N$-labeled "Z-domain" of the 71-residue *Staphylococcal* protein A (12) in 90% H_2O/10% D_2O (20 mM K-PO_4; pH, 6.5). The protein's overall rotational correlation time obtained from polypeptide backbone ^{15}N $T_{1\rho}/T_1$ ratios (13) is 4.5 ns—i.e., within the range encountered for proteins from about 5 to 15 kDa.

Nine RD NMR experiments were used in conjunction with 3D HNNCACB (4–8, 14–17). Fig. 1 surveys their names and the correlated chemical shifts. In the nomenclature of the RD NMR experiments underlined letters indicate chemical shifts obtained in a common dimension. Fig. 2 displays peak patterns observed in the projected dimensions, and Fig. 6 (which is published as supporting information on the PNAS web site, www.pnas.org)

Abbreviations: HTP, high-throughput; RD, reduced-dimensionality; 2D, 3D, 4D, two-, three-, four-dimensional; COSY, correlation spectroscopy; TOCSY, total correlation spectroscopy; TR, triple-resonance; S/N, signal-to-noise.

†To whom reprint requests should be addressed at: Department of Chemistry, State University of New York, Buffalo, NY 14260. E-mail: szypersk@chem.buffalo.edu.

‡Present address: Center for Advanced Research for Biotechnology, University of Maryland, Rockville, MD 20850.

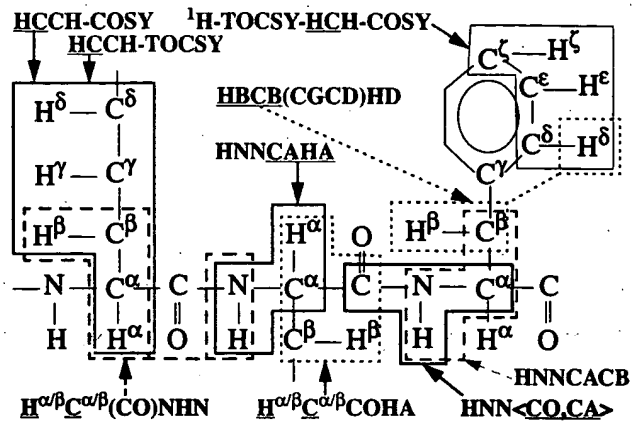


Fig. 1. Polypeptide chemical shifts correlated by the spectra constituting the standard set of experiments for RD NMR-based protein resonance assignment. The nuclei for which the chemical shifts are obtained from a given experiment are boxed and labeled accordingly. Experiments providing sequential and intraresidue connectivities are in black and gray, respectively.

depicts the magnetization transfer pathways. Three-dimensional $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN (6, 7) and HACA(CO)NHN yield *sequential* connectivities. Three-dimensional HNNCAHA (7), HNNCACB, and $H^{\alpha/\beta}C^{\alpha/\beta}$COHA provide *intra*residue connectivities, and 3D HNN⟨CO,CA⟩ (5) offers both intraresidue $^1H^N-^{13}C^\alpha$ and sequential $^1H^N-^{13}C'$ connectivities. Three-dimensional HCCH-correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY) provide assignments of *aliphatic* side chains, whereas two-dimensional (2D) HBCB(CGCD)HD and ^{1}H-TOCSY-relayed HCH-COSY provide those of the *aromatic* spins. The NMR pulse schemes of the hitherto unpublished RD NMR experiments are provided in Figs. 7–12, which are published as supporting information on the PNAS web site. The maximal evolution times, as well as the resulting measurement times are given in Table 1. RD NMR experiments in which ^{1}H and ^{13}C are simultaneously observed in a projected dimension were acquired with virtually the same maximal evolution times in $t_1(^{13}C/^1H)$ to enable accurate matching of peak patterns (Fig. 2). For HNNCAHA and HNN⟨CO,CA⟩, central peaks were derived from incomplete polarization transfer (6, 7). For others, ^{13}C magnetization present at the end of the refocusing period of the initial polarization transfer from ^{1}H to ^{13}C was recruited, which yields two subspectra containing the peak pairs and central peaks, respectively (Fig. 2). In view of potential peak overlap, proper setting of the radio-frequency (rf) carriers is crucial. In HNNCAHA, for example, this allows one to place central peaks, and up-field and down-field component of the peak pairs into separate spectral regions (7). When data are collected in this manner, peak overlap does not increase when compared with HNNCA.

The relative sensitivity of NMR experiments was analyzed by determining the yield of peak detection—i.e., the ratio of observed peaks over the total number of expected peaks—and by separately assessing the S/N ratio distributions of peaks belonging to either peak pairs or central peaks. To rank the experiments (Table 1) according to sensitivity, only peaks encoding the prime information of a given spectrum were considered—e.g., *intra*residue connectivities in HNNCAHA and HNNCACB, correlation peaks in HCCH-COSY and *relay* connectivities in HCCH-TOCSY. The averaged S/N ratios were divided by the square root of the measurement time (Table 1) and scaled relative to the most sensitive experiment yielding peak pairs—i.e., HACA(CO)NHN. To avoid a bias from longer transverse relaxation times in several flexibly disordered terminal residues, the N-terminal octapeptide segment comprising residues "−13" to "−6" (in the numbering chosen in ref. 12) was not considered in these sensitivity analyses.

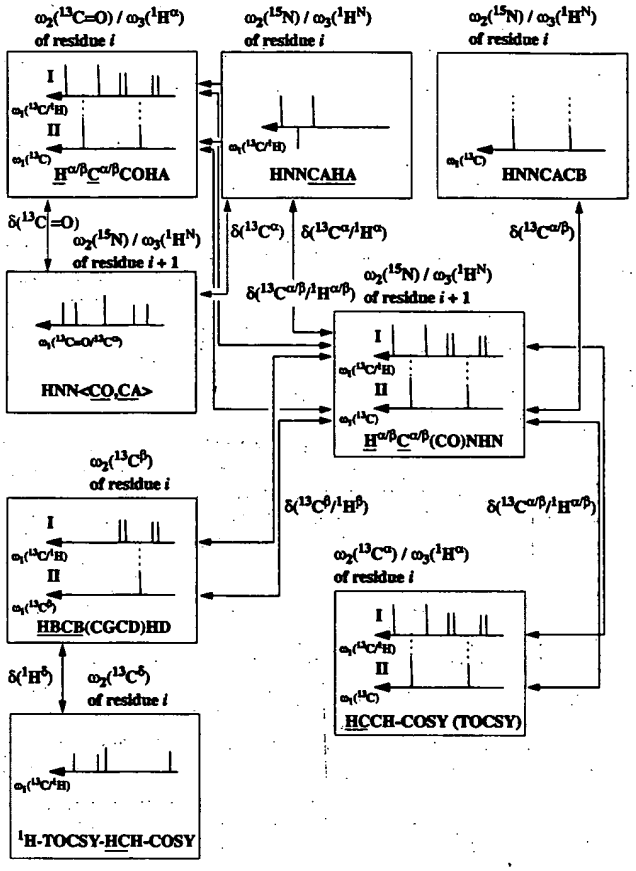


Fig. 2. Schematic presentation of the RD NMR-based HTP resonance assignment strategy by using the standard set of experiments of Fig. 1. The central role of 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN is shown for establishing sequential connectivities via (*i*) $^{13}C^\alpha$ and $^1H^\alpha$ shift measurements (HNNCAHA, Fig. 3), (*ii*) $^{13}C^\alpha$ and $^{13}C^\beta$ shift measurements (HNNCACB), and (*iii*) ^{13}C=O shift measurements ($H^{\alpha/\beta}C^{\alpha/\beta}$COHA/HNNCAHA and HNN⟨CO,CA⟩; see supporting information), as well as for assigning aliphatic (HCCH-COSY/TOCSY; Fig. 4) and aromatic side chains (HBCB(CGCD)HD and ^{1}H-TOCSY-relayed HCH-COSY; Fig. 5). Black double-headed arrows indicate connectivities that are established based on matching of peak patterns along $\omega_1(^{13}C/^1H)$. Gray arrows indicate that the combined use of the two spectra requires the conversion of in-phase splittings into chemical shifts. Each box shows the peak patterns expected along ω_1, and the chemical shifts that are measured in the other dimensions are given above the corresponding boxes. Two cross sections are sketched for RD NMR experiments yielding two subspectra I and II, which comprise peak pairs and central peaks, respectively (6).

The program AUTOASSIGN (V. 1.7.3; refs. 10 and 11) was extended for analysis of RD TR NMR experiments. The input included 3D peak lists derived from 3D HACA(CO)NHN, $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN and HNNCAHA, a 3D HNNCO peak list from 3D HNN⟨CO,CA⟩ and the 3D HNNCACB peak list. Details of the computational protocol are available in *Supporting Text*, which is published as supporting information on the PNAS web site. The AUTOASSIGN program, the input RD NMR peak lists, and the output resonance assignment lists are available from the authors.

Results and Discussion

The data collection times and the relative sensitivities of the experiments are shown in Table 1. For those providing sequential connectivities—i.e., 3D HACA(CO)NHN,

Table 1. Multidimensional NMR experiments

Experiment (acquisition of central peaks)*	Indirect† dimension(s)	t_{max}, ms; complex points	Measurement time, h	Relative sensitivity (peak pairs/ central peaks)
Sequential backbone connectivities (3D spectra)				
$H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN	$\omega_1(^{13}C^{\alpha/\beta}/^{1}H^{\alpha/\beta})$	6.3; 95	9.2	(0.56/0.34)
(^{13}C)	$\omega_2(^{15}N)$	21.5; 28		
HACA(CO)NHN	$\omega_1(^{13}C^{\alpha}/^{1}H^{\alpha})$	6.5; 54	5.4	(1.00‡/0.81)
(^{13}C)	$\omega_2(^{15}N)$	21.5; 28		
Intraresidual backbone connectivities (3D spectra)				
HNNCAHA	$\omega_1(^{13}C^{\alpha}/^{1}H^{\alpha})$	6.6; 51	5.0	(0.41/0.27)
(INEPT)	$\omega_2(^{15}N)$	21.5; 28		
$H^{\alpha/\beta}C^{\alpha/\beta}$COHA	$\omega_1(^{13}C^{\alpha/\beta}/^{1}H^{\alpha/\beta})$	6.3; 95	10.0	(0.22/0.11)
(^{13}C)	$\omega_2(^{13}C{=}O)$	17.8; 32		
HNNCACB	$\omega_1(^{13}C^{\alpha/\beta})$	6.6; 56	8.0	(0.56)
	$\omega_2(^{15}N)$	21.5; 28		
Intra- and sequential-backbone connectivities (3D spectrum)				
HNN⟨CO,CA⟩	$\omega_1(^{13}C^{\alpha}/^{13}C{=}O)$	8.0/16.0§; 54	5.5	(0.54/1.41)
(INEPT)	$\omega_2(^{15}N)$	21.5; 28		
Assignment of aliphatic resonances (3D spectra)				
HCCH-COSY	$\omega_1(^{13}C/^{1}H)$	6.3; 95	6.2	(0.34/0.25)
(^{13}C)	$\omega_2(^{13}C)$	6.4; 20		
HCCH-TOCSY¶	$\omega_1(^{13}C/^{1}H)$	6.3; 95	7.0	(0.19/n.d.)
(^{13}C)	$\omega_2(^{13}C)$	6.4; 20		
Assignment of aromatic resonances (2D spectra)				
HBCB(CGCD)HD	$\omega_1(^{13}C/^{1}H)$	6.3; 95	5.3	(0.45/0.33)
(^{13}C)				
^{1}H-TOCSY-HCH-COSY‖	$\omega_1(^{13}C/^{1}H)$	15; 150	3.4	(0.76/–)

One millimolar solution of "Z-domain" of *Staphylococcal* protein A at T = 25°C. The radio-frequency (rf) carrier for ^{1}H-frequency labeling in the projected "HC"-dimensions in which the chemical shifts of the aliphatic moieties are measured was set to 0 ppm. In 2D ^{1}H-TOCSY-HCH-COSY, the ^{1}H rf carrier was set to the position of the water line throughout. t_{max} denotes the maximal evolution time. The suite of experiments in this table can provide complete resonance assignments of proteins, excluding only the side chain NH_n moieties, the CH^{ε} groups of histidinyl, and the $CH^{\varepsilon 3}$, $CH^{\zeta 2,3}$, and $CH^{\eta 2}$ groups of tryptophanyl residues (which can be obtained as described in ref. 17). Notably, Z-domain does not contain tryptophans.

*Approach 1: Use of incomplete polarization transfer (rows labeled with "INEPT"); Approach 2: use of ^{13}C steady state magnetization (rows labeled with "^{13}C").

†Direct dimension: t_{max} = 73 ms/512 complex points.

‡The average signal-to-noise (S/N) ratio of peaks observed in this subspectrum was 33.2.

§The increment for $^{13}C^{\alpha}$ chemical shift evolution was scaled by a factor of 0.5 relative to the values used to sample $^{13}C{=}O$ evolution (5).

¶The mixing times for the ^{13}C-TOCSY relay was set to 21 ms. The S/N ratios for the double-relay central peaks were too low to be accurately evaluated.

‖The mixing time for the ^{1}H-TOCSY relay was set to 25 ms. The acquisition of central peaks is prevented by the use of spin-lock purge pulses (flanking the total correlation relay) to obtain pure phases.

$H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, and HNN⟨CO,CA⟩ (Fig. 1)—peak pair and central peak detection was complete (100%). Among the experiments providing *intra*residue connectivities, both 3D HNNCAHA and 3D HNNCACB likewise exhibited 100% yield, whereas a few were missing in $H^{\alpha/\beta}C^{\alpha/\beta}$COHA (yield, 98% of peak pairs; 91% of central peaks). In part, this is because magnetization is detected on the $^{1}H^{\alpha}$ protons close to the water resonance. Peak pair detection in 3D HCCH-COSY was nearly complete (90% of peak pairs, 93% of central peaks), whereas the yield of *relayed COSY* peak pairs in 3D HCCH-TOCSY was slightly lower (81%). To some extent this was due to signal overlap and not the lack of sensitivity. Nearly complete signal detection was also observed in 2D HBCB(CGCD)HD (100%) and ^{1}H-TOCSY-relayed HCH-COSY (90% of peak pairs). The detailed S/N analysis revealed (*i*) outstanding sensitivity for detection of peak pairs in 3D HACA(CO)NHN, (*ii*) about similar sensitivity for 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, HNNCAHA, HNN⟨CO,CA⟩, HNNCACB (nowadays routinely used up to around 25 kDa), HCCH-COSY, and 2D ^{1}H-TOCSY-relayed HCH-COSY, and (*iii*) reduced sensitivity for 3D $H^{\alpha/\beta}C^{\alpha/\beta}$COHA, 2D HBCB(CGCD)HD, and relay peak detection in 3D HCCH-TOCSY. (Note that analysis of the spectra for assignment of the aromatic spin systems was somewhat impeded by the small number of aromatic residues in Z-domain, as well as their partially flexibly disordered nature.) In accordance with the very high sensitivity of HNNCO, central peak detection in HNN⟨CO,CA⟩ is by far the most sensitive and serves for secure spin system identification in cases of overlap in 2D [^{15}N,^{1}H]-heteronuclear sequential quantum correlation (HSQC) (5, 10).

The analysis summarized in Table 1 allows one to devise a strategy for RD NMR-based HTP resonance assignment in which 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN establishes sequential backbone connectivities and connectivities to both the aliphatic and aromatic side chains (Fig. 2). Firstly, the peak patterns along $\omega_1(^{13}C^{\alpha/\beta}/^{1}H^{\alpha/\beta})$ in subspectra I and II of 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN enable sequential resonance assignment in combination with HNNCAHA (Fig. 3) and HNNCACB, respectively. Complementary recording of 3D $H^{\alpha/\beta}C^{\alpha/\beta}$COHA and HNN⟨CO,CA⟩ contributes polypeptide backbone $^{13}C'$ chemical shift measurements for establishing sequential assignments: the intraresidue

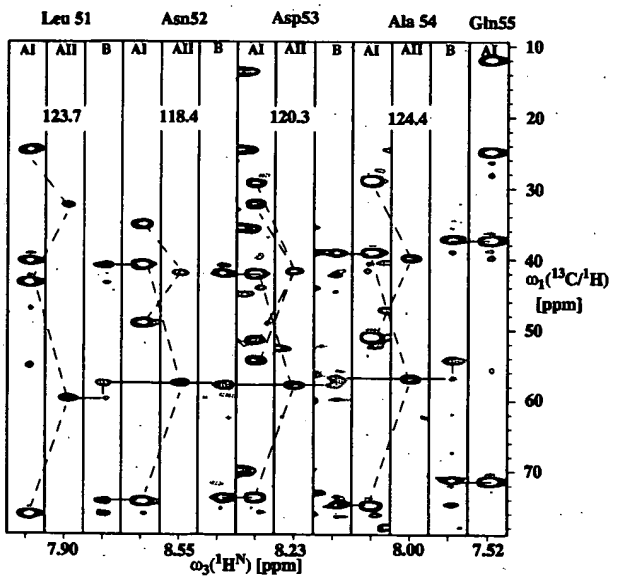


Fig. 3. Sequential resonance assignment from 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN/3D HN-NCAHA. Contour plots of [$\omega_1(^{13}C)$, $\omega_3(^1H^N)$]-strips taken from subspectrum I (strips labeled with AI) and subspectrum II (strips labeled with AII) of 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN, and from 3D HNN<u>CAHA</u> (strips labeled with B) are shown. The strips were taken at the ^{15}N chemical shifts (indicated at the top) of residues 51 to 55 and are centered about their $^1H^N$ chemical shift. The sequence-specific resonance assignments of the amide chemical shifts are given at the top of each strip and are referred to as *i*. $\Omega(^1H^{\alpha/\beta}{}_{i-1})$ and $\Omega(^{13}C^{\alpha/\beta}{}_{i-1})$ obtained from 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN are given in the strips AI and AII of residue *i*. Corresponding peak pairs in AI and central peaks in AII are connected by dashed lines, and sequential connectivities are indicated by solid lines for both peak pairs and central peaks. Dashed and solid contour lines represent negative and positive peaks, respectively, and sequential connectivities established via the central peaks and via the peak pairs are indicated by solid and dotted lines, respectively. Note that the near-degeneracy of $^{13}C^\alpha$ chemical shifts in the polypeptide segment Asn-52–Asp-53–Ala-54 is neatly resolved by the measurement of $^1H^\alpha$ chemical shifts encoded in the in-phase splittings of the peak pairs. Chemical shifts are relative to 2,2-dimethyl-2-silapentane-5-sulfonate (DSS).

$^{13}C'$ correlation is obtained by $\omega_1(^{13}C^{\alpha/\beta}/^1H^{\alpha/\beta})$ peak pattern matching of $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN with $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$COHA, and the sequential $^{13}C'$ correlation is inferred from (13)C^α, ^{15}N, and $^1H^N$ chemical shifts in HNN(<u>CO,CA</u>) (see Fig. 7). Secondly, comparison of $\omega_1(^{13}C^{\alpha/\beta}/^1H^{\alpha/\beta})$ peak patterns with 3D <u>HCCH</u>-COSY and -TOCSY connects the $C^{\alpha/\beta}/H^{\alpha/\beta}$ chemical shifts with those of the more peripheral aliphatic side chain spins (Fig. 4), whereas comparison of $\omega_1(^{13}C^\beta/^1H^\beta)$ peaks with 2D <u>HBCB</u>-(CGCD)HD and subsequent linking with $^1H^\delta$ chemical shifts detected in 2D ^{1}H-TOCSY-relayed <u>HCH</u>-COSY affords assignment of the aromatic spin systems (Fig. 5). For many residues the two β-proton chemical shifts are nondegenerate, and the connection of $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN with <u>HBCB</u>(CGCD)HD or <u>HCCH</u>-COSY/TOCSY may then rely on comparison of the three chemical shifts of $^1H^{\beta2}$, $^1H^{\beta3}$, and $^{13}C^\beta$. In general, the identification of peaks pairs is complicated when chemical shift degeneracy in the other dimension occurs, but central peak acquisition (Figs. 2 and 3) addresses this complication in a straightforward fashion (6). Importantly, however, pairs of peaks generated by a chemical shift in-phase splitting have quite similar intensity. Usually this does not hold for two arbitrarily selected peaks, because the nuclear spin relaxation times vary within each residue as well as along the polypeptide chain. The peak pairs are thus "labeled" with the relaxation times, making the pair identification obvious in most cases (Figs. 3–5). This is also advantageous for automated peak picking.

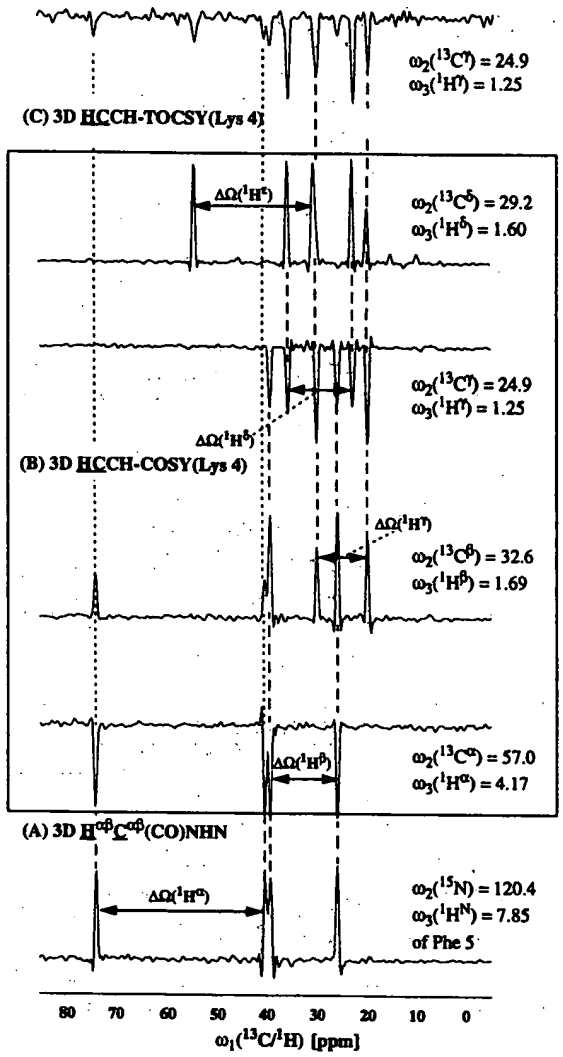


Fig. 4. Assignment of aliphatic spin systems exemplified for Lys-4. Cross sections were taken along $\omega_1(^{13}C/^1H)$ from the subspectra I comprising peak pairs of (*A*) 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN, (*B*) 3D <u>HCCH</u>-COSY, and (*C*) 3D <u>HCCH</u>-TOCSY. The signals in 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN were detected on the backbone amide proton of the succeeding residue Phe 5 (the ^{15}N and $^1H^N$ chemical shifts are indicated on the right). The cross sections taken from <u>HCCH</u>-COSY exhibit signals that were detected on $^1H^\alpha$, $^1H^\beta$, $^1H^\gamma$, and $^1H^\delta$ of Lys-4, respectively (from the bottom to the top), and the corresponding chemical shifts are given on the right. The in-phase splittings encode the $^1H^\beta$, $^1H^\gamma$, $^1H^\delta$, and $^1H^\varepsilon$ chemical shifts and serve to obtain the desired correlations (indicated by dashed vertical lines). Note that the peak signs vary because of aliasing along $\omega_2(^{13}C)$. In *C*, the signal was detected on $^1H^\gamma$ of Lys-4, and the crucial $^\alpha$CH–$^\gamma$CH-TOCSY *relay* connectivity is indicated. Chemical shifts are relative to 2,2-dimethyl-2-silapentane-5-sulfonate (DSS).

The suite of the nine experiments outlined in Fig. 1 complemented by 3D <u>HACA</u>(CO)NHN forms a "standard set" for RD NMR-based protein resonance assignment. For Z-domain, the entire set was recorded within only 65 h by using a conventional TR NMR probe at 600 MHz (Table 1). The high redundancy of these seven projected 4D, one conventional 3D, and two projected 3D spectra provides a very efficient resonance assignment strategy, which also profits from the fact that the detection of symmetric RD NMR peak pairs greatly facilitates the identification of peaks close to the noise level. Importantly, the information provided by projected 4D spectra cannot be obtained by recording twice the number of 3D spectra: in cases of chemical shift degeneracy a chemical shift quartuple is not equivalent to two shift triples.

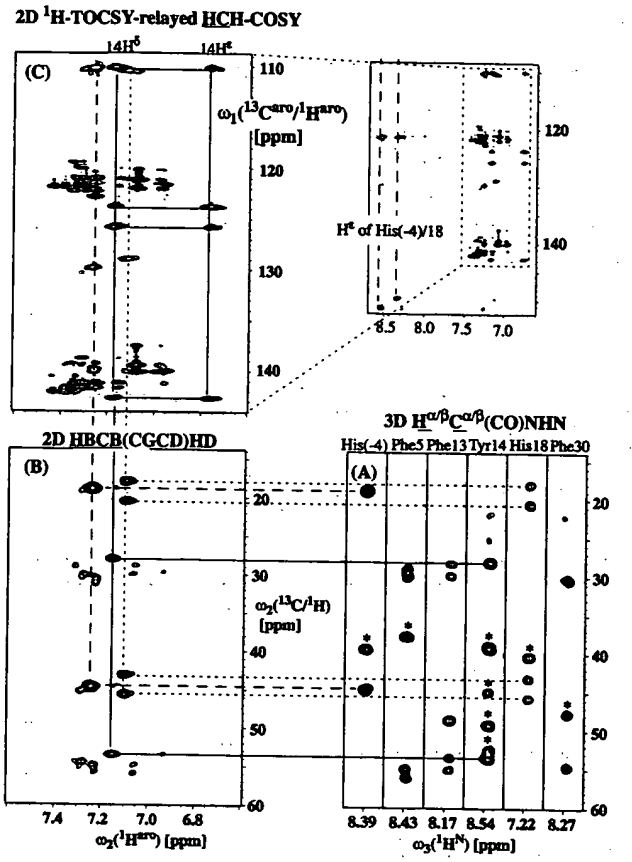


Fig. 5. Assignment of aromatic side chains exemplified for His(-4) and His-18, and Tyr-14. A composite plot of [$\omega_1(^{13}C/^1H)$,$\omega_3(^1H^N)$]-strips taken from 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN comprising the $\omega_1(^{13}C/^1H)$ peaks of all aromatic side chains in the polypeptide segment −5 to 58 of Z-domain, the 2D HBCB(CGCD)HD spectrum (*B*), and a spectral region taken from 2D ^{1}H-TOCSY-relayed HCH-COSY (*C*) are shown. The entire 2D ^{1}H-TOCSY-relayed HCH-COSY spectrum, which also contains cross peaks arising from $^\varepsilon$CH of the histidinyl residues, is shown in the upper right of the figure. Correlations belonging to His(-4) and His-18, and Tyr-14 are connected with long-dashed, dashed, and gray solid lines, respectively. In *C*, peaks arising from $^\alpha$CH moieties (which are not required for connecting the aromatic spin systems) are labeled with an asterisk. Chemical shifts are relative to 2,2-dimethyl-2-silapentane-5-sulfonate (DSS).

It is of significant practical advantage that the sensitivity within the standard set varies only by about a factor of two (Table 1). This facilitates the prediction of minimal required measurement times (roughly a multiple of the measurement time of a single experiment). In fact, the S/N ratios observed in the first experiment allow one to adjust measurement times while data acquisition is in progress. For Z-domain, six RD NMR experiments were actually sufficient to provide complete backbone and side chain resonance assignments (3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, HNNCAHA, HCCH-COSY/TOCSY, 2D HBCB(CGCD)HD, and ^{1}H-TOCSY-relayed HCH-COSY recorded in 36 h; Table 1), and those can be considered as a "minimal set" for proteins up to around 10 kDa. As expected, chemical shifts agreed well with those previously obtained at 30°C by using conventional TR NMR (12).

For proteins above ≈15 kDa, recording of highly sensitive 3D HACA(CO)NHN promises (*i*) to yield spin systems that escape detection in $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, and (*ii*) to offer the efficient distinction of α- and β-moiety resonances by comparison with $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN. Moreover, Nietlispach *et al.* (18) have recorded 4D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN for 50% random fractionally deuterated proteins reorienting with correlation times up to around 20 ns (corresponding to ≈20–30 kDa at ambient temperature). Thus, the 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, or a transverse relaxation optimized version thereof (19), may well maintain the role outlined in Fig. 2 for assigning partially deuterated proteins at least up to about that size, in particular when using cryogenic probes (20).

Cryogenic probes reduce measurement times by about a factor of 10 or more (21). In the sensitivity-limited regime, where sampling requirements do not provide a bound for the minimal measurement time, the required signal could have been recorded for the standard set of ten experiments in about 6.5 h (Table 1). Using the current implementations (Table 1), the standard set could have been recorded with a single transient per increment and without central peak detection, considering the "spin relaxation time labeling" of peak pairs. This would then reduce the total measurement time to about 18 h. Hence, RD NMR promises to allow a rather close adjustment of measurement times to sensitivity requirements.

HTP employment of RD NMR requires strong computer support for data analysis. For the backbone and $^{13}C^\beta$ resonances, this issue was addressed by extending the program AUTOASSIGN (10, 11) to analyze 3D HACA(CO)NHN, $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, HNNCAHA, HNN(CO,CA), and HNNCACB (recorded in 33 h; Table 1). AUTOASSIGN determined 94.5% (345 of 366) of all backbone $^1H^N$, ^{15}N, ^{13}C = O, $^{13}C^\alpha$, $^{13}C^\beta$, and $^1H^\alpha$ chemical shifts with an error rate of only 0.9%, and obtained at least three chemical shifts for 61 of 63 residues (97%). Evidently, the good spectral resolution of the RD spectra (Table 1) greatly supported the efficient automated analysis. Importantly, the chemical shifts rapidly obtained from AUTOASSIGN directly revealed the protein's secondary structure (22): the helical boundaries derived independently from this chemical shift analysis and the NMR solution structure (12) are virtually identical. The extension of AUTOASSIGN for assigning side chain chemical shifts is currently in progress. This will allow one to automatically obtain nearly complete resonance assignments of proteins.

Conclusion

RD NMR is a powerful approach to avoid the "sampling limited acquisition regime." This is of outstanding interest in view of the forthcoming era of cryogenic probes. In particular, the resulting rapid determination of protein secondary structure from chemical shifts (22) will greatly support fold prediction (23), protein target selection, and construct optimization in structural genomics. Considering that (*i*) sensitivity and sweep widths increase with increasing magnetic fields and that (*ii*) the widespread use of cryogenic probes will greatly boost the sensitivity of our spectrometers, we expect a "change in paradigm" in biological NMR spectroscopy with a new focus on research addressing the caveat of sampling limitation. Data processing protocols reducing the number of data points in the indirect dimensions without sacrificing spectral resolution, such as linear prediction and maximum entropy methods (24, 25), nonlinear sampling (25, 26), and possibly the recently introduced filter diagonalization method (27, 28), appear to be of keen interest to further enhance the impact of RD NMR.

This work was supported by a State University of New York start-up fund (to T.S.), National Science Foundation Grants MCB 0075773 (to T.S.) and DBI-9974200 (to H.N.B.M.), and National Institutes of Health (Northeast Structural Genomics Consortium) Grant P50 GM62413-01 (to T.S. and G.T.M.).

1. Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C. & Szyperski, T. (2000) *Nat. Struct. Biol.* 7, 982–984.

2. Moseley, H. N. B. & Montelione, G. T. (2000) *Curr. Opin. Struct. Biol.* 9, 635–642.

3. Cavanagh, J., Fairbrother, W. J., Palmer, A. G. & Skelton, N. J. (1996) *Protein NMR Spectroscopy* (Academic, San Diego).
4. Szyperski, T., Wider, G., Bushweller, J. H. & Wüthrich, K. (1993) *J. Am. Chem. Soc.* **115,** 9307–9308.
5. Szyperski, T., Braun, D., Fernández, C., Bartels, C. & Wüthrich, K. (1995) *J. Magn. Reson. B* **108,** 197–203.
6. Szyperski, T., Braun, D., Banecki, B. & Wüthrich, K. (1996) *J. Am. Chem. Soc.* **118,** 8146–8147.
7. Szyperski, T., Banecki, B., Braun, D. & Glaser, R. W. (1998) *J. Biomol. NMR* **11,** 387–405.
8. Szyperski, T., Wider, G., Bushweller, J. H. & Wüthrich, K. (1993) *J. Biomol. NMR* **3,** 127–132.
9. Reif, B., Hennig, M. & Griesinger, C. (1997) *Science* **276,** 1230–1233.
10. Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, A., Chien, C.-Y., Powers, R. & Montelione, G. T. (1997) *J. Mol. Biol.* **269,** 592–610.
11. Moseley, H. N. B., Monleon, D. & Montelione, G. T. (2001) *Methods Enzymol.* **339,** 91–108.
12. Tashiro, M., Tejero, R., Zimmerman, D. E., Celda, B., Nilsson, B. & Montelione, G. T. (1997) *J. Mol. Biol.* **272,** 573–590.
13. Szyperski, T., Luginbühl, P., Otting, G., Güntert, P. & Wüthrich, K. (1993) *J. Biomol. NMR* **3,** 151–164.
14. Wittekind, M. & Müller, L. (1993) *J. Magn. Reson. B* **101,** 201–205.
15. Kay, L. E. (1993) *J. Am. Chem. Soc.* **115,** 2055–2057.
16. Yamazaki, T., Forman-Kay, J. D. & Kay, L. E. (1993) *J. Am. Chem. Soc.* **115,** 11054–11055.
17. Zerbe, O., Szyperski, T., Ottiger, M. & Wüthrich, K. (1996) *J. Biomol. NMR* **7,** 99–106.
18. Nietlispach, D., Clowes, R. T., Broadhurst, R. W., Ito, Y., Keeler, J., Kelly, M., Ashurst, J., Oschkinat, H., Domaille, P. J. & Laue, E. D. (1996) *J. Am. Chem. Soc.* **118,** 407–415.
19. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 12366–12371.
20. Styles, P., Soffe, N. F., Scott, C. A., Cragg, D. A., White, D. J. & White, P. C. (1984) *J. Magn. Reson.* **60,** 397–404.
21. Monleon, D., Colson, K., Moseley, H. N. B., Anklin, C., Oswald, R., Szyperski, T. & Montelione, G. T. (2002) *J. Struct. Funct. Genom.*, in press.
22. Wishart, D. S., Sykes, B. D. & Richards, F. M. (1992) *Biochemistry* **18,** 1647–1651.
23. Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998) *J. Mol. Biol.* **277,** 419–448.
24. Stephenson, D. S. (1988) *Prog. NMR Spectroscopy* **20,** 515–626.
25. Hoch, J. C. & Stern, A. S. (1996) *NMR Data Processing* (Wiley–Liss, New York).
26. Schmieder, P., Stern, A. S., Wagner, G. & Hoch, J. C. (1994) *J. Biomol. NMR* **4,** 483–490.
27. Wall, M. R. & Neuhauser, D. (1995) *J. Chem. Phys.* **112,** 8011–8022.
28. Hu, H., Van, Q. N., Mandelshtam, V. A. & Shaka, A. J. (1998) *J. Magn. Reson.* **134,** 76–87.

EXHIBIT 7

Proceedings of the National Academy of Sciences
Fax Cover Sheet

DATE: April 19, 2002

TO: Dr. Thomas Szyperski
Dept of Chemistry
University at Buffalo
The State University of New York
Buffalo, NY 14260

Phone: 716-645-6800, ext. 2245 **Fax:** 716-645-7338
Email: szypersk@chem.buffalo.edu

FROM: PNAS Office **Re:** Ms. No. 02-2245
1055 Thomas Jefferson Street, NW
Room 2013
Washington, DC 20007
Phone: (202) 334-2679 **Fax:** (202) 334-2739
Email: pnas@nas.edu **Internet:** http://www.pnas.org

Number of pages including cover sheet: 1

Author reference: Szyperski et al.
Title: Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment

We are pleased to inform you that the PNAS Editorial Board has given final approval of your paper for publication in PNAS. You will receive reprint order information and an invoice for publication fees with your proofs. Please direct all production questions to Cadmus Professional Communications: Eric Sarlin (phone: 410-691-6997; e-mail: sarline@cadmus.com) or Angela Hughes (phone: 410-691-6404; e-mail; hughesa@cadmus.com); toll-free number: 800-257-5529; fax: 410-691-6220.

As of late January 2000, PNAS papers may be published online before print at www.pnas.org in a new feature, "PNAS Early Edition," which will appear every Tuesday. Papers may be published online one to four weeks before they appear in print. Authors who return proofs quickly and keep changes to a minimum will get maximum speed of publication. As the system progresses, we will go to daily online publication. The date a paper appears online in PNAS Early Edition is the publication date of record and will be posted with the article text online.

When a paper is accepted for publication, it is under embargo and not for public release before 5 p.m. Eastern time, the day before publication. Authors may talk with the press about their work, but should coordinate this with the NAS Office of News and Public Information (ONPI), so that reporters are aware of PNAS policy and understand that papers are embargoed until the date of publication. Please refer reporters to ONPI at 202-334-2138.

Authors are invited to submit scientifically interesting and visually arresting cover illustrations. Send two glossy prints 20 cm wide by 12 cm high, with the top indicated, and a brief lay language caption (50-60 words) to the PNAS office address above (phone: 202-334-2679) as soon as possible. Label the submission with manuscript number, author name, phone, fax, and e-mail. Digital art may accompany prints (color art must be in EPS format and CMYK mode). See the Information for Authors for details.

PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, U.S.A.

Request for opinion on manuscript by Reviewer # II

Thomas Szyperski, Deok C. Yeh, Dinesh K. Sukumaran, Hunter N.B. Moseley, and Gaetano T. Montelione

Title Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment

The *Proceedings of the National Academy of Sciences, U.S.A.*, a multidisciplinary journal, publishes brief reports of original research of exceptional importance and novelty. The Academy Member listed at the bottom of this form is asking your opinion on the following points, together with any other comments you may offer. Please reply to all questions. The Editorial Board considers the first two the most important. Please note that the Editorial Policy is that the referees should remain anonymous.

1. Is the overall quality of this paper suitable for this journal? "Yes" implies that the paper is in the top 10% in its field.
   ☑ Yes ☐ No ☐ Don't know
2. Is this contribution of sufficient general interest to justify publication in the *Proceedings* rather than a specialty journal?
   ☑ Yes ☐ No ☐ Don't know
3. Does the evidence justify the conclusions drawn? ☑ Yes ☐ No ☐ Don't know
4. Is this paper clearly written for a diverse audience of scientists? ☑ Yes ☐ No ☐ Don't know
5. Are the procedures described sufficiently well that the work can be repeated? ☑ Yes ☐ No ☐ Not relevant
6. Comments (use additional pages if necessary). If the answers to questions 1 and 2 are Yes, please describe here the aspects of this paper that are novel and important:

This paper describes the generalization of the concept of "reduced dimensionality" and its application to the NMR pulse sequences for protein resonance assignment. Efficiency in NMR data collection is one of the important concerns especially in structural genomics projects. This is not only a proposal of a concrete answer but also a proposal suggestive to other general problems.

However, I recommend to change the following subtle points.

1. The meaning of underlines drawn under the names of pulse sequences should be given.
2. The six pulse sequences (p4, 11) that are newly implemented should be described at least in the supporting information section.
3. The number of experiments listed as a standard set is 10, but the number given in the text is 9 (p4, 111-19). HACA(CO)NHN should be counted.
4. The statement that the relaxation time for 13C is shorter than that of 1H(p5, 12) is wrong, because 1H magnetization recovers faster through cross relaxation from CH3 groups.
5. If the possibility of automated assignment of the side chain signals would be described, it must be more interesting.

7. If the manuscript is revised, I would be prepared to rereview it. ☑

Please return original within two weeks to the Member who asked you to referee the paper.

Dr. Herbert A. Hauptman
Name of Member

Hauptman-Woodward Medical Research Inst., Inc.
Address of Member

73 High Street

Buffalo, New York 14203-1196

The reduced dimensionality experiments which are discussed in this paper are very important for enhancing information content to enable automated assignment of NMR resonances, and thereby high throughput structure determination by NMR. The concept of reduced dimensionality experiments is not new, but this manuscript puts together (for the first time to the knowledge of this reviewer) a suite of experiments, which are specifically useful for automated data analysis. The particular set of experiments seems to have been selected to be compatible with the AUTOASSIGN program from one of the authors, other sets may be optimal in other contexts but what is presented demonstrates the concept very well.

A minor weakness of this paper is that it presents the concepts in NMR jargon, which will likely limit readers to a fairly specialized audience. Even realizing that space for PNAS papers is usually at a premium I would recommend an expanded introductory section making the ideas more accessible to a broader audience – stressing more the concept of a broad experimental optimization, in terms of information content and time, that can be accomplished with available technology.

A few specific suggestions to clarify various points:

The number of amino acid residues in the "Z-domain" of protein A should be given.

In describing experiments it would be worthwhile pointing out that the underlined spins are those which coevolve to reduce dimensionality.

I strongly recommend putting the pulse sequences for these reduced dimensionality experiments into the supporting material available on the web rather than having them only "available from the authors".

At the top of page 5 the statement "^{13}C magnetization was recruited" is unclear, further reading explains but it would be better to give a clearer initial description.

At the bottom of page 6 the meaning of "plays a central role" is not really clear, one would think that in automated analysis of data all experiments are optimally utilized and so none really plays a central role. The discussion following this statement really sounds more like a manual rather than automated assignment method.

At the top of page 9 the statement about time is confusing, it is first stated that the ten experiments of Table 1 could have been recorded in 6 hours, but then states that even with one transient it would take 17 hours. The 60 hrs / 10 is meaningless since the experiments could not really be done in this time, just state that the realistic minimum time is 17 hrs.

EXHIBIT 8

NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230
DIVISION OF MOLECULAR AND CELLULAR BIOSCIENCES

Dr. Thomas A. Szyperski
Department of Chemistry
SUNY at Buffalo
Amherst-Campus
816 Natural Sciences Complex
Buffalo, NY 14260

Ref: MCB-9983995

Dear Dr. Szyperski:

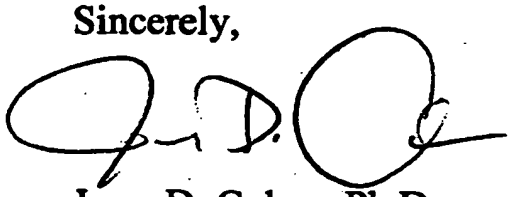
We regret to inform you that the National Science Foundation is unable to support your proposal entitled "CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics".

A number of factors are considered in evaluating each proposal submitted to the Foundation. Of these, scientific merit is most important. Other factors include the relation of the proposed research to other research in the area and the distribution of funds among the various areas of biological sciences. Many meritorious proposals cannot be funded simply because of the limited resources available.

For further information concerning the evaluation of your proposal, please contact the Program Officer whose name and telephone number appears on the enclosed *Context Statement.* Copies of the reviews of your proposal and a *Panel Summary,* outlining the salient points raised in the panel discussion of your proposal, are enclosed. These are for your personal use and are not made available by the Foundation to anyone else. They may be helpful to you in preparing future proposals.

Although we were unable to support this proposal, we would be pleased to consider any future proposals you may wish to submit.

Sincerely,



Jerry D. Cohen, Ph.D.
Acting Deputy Division Director

Panel Summary
Molecular Biophysics
Szyperski, Thomas
MCB-9983995

While recognizing that the PI pioneered the RD method in NMR, the panel believed that the PI must address important issues raised by the reviewers. In particular, the panel questioned whether the RD method will be broadly applicable, particularly to larger proteins with congested spectra. In addition, the PI should address the extent to which RD approaches truly reduce the main bottlenecks in NMR structure determination for high throughput structural genomics. The PI should address in more detail possible pitfalls in the proposed approaches and suggest alternatives (what methods other than RD might be used in structural genomics for proteins for which RD is not successful).The panel viewed the education plan and collaborations as strong aspects of the proposal. The panel believes that the investigator has high potential and encourages a suitable revision of this proposal.

Overall Rating: Good

PROPOSAL NO.: 9983995
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Good

REVIEW:
This is a CAREER proposal directed towards (1) developing reduced dimentionality NMR experiments for solution structure elucidation and (2) developing Bio-NMR and structural biology courses at SUNY Buffalo. Dr. Szyperski will develop new experiments within the RD approach- for instance one approach will allow measurement of residual dipolar couplings in liquid crystalline media for use in structure calculations. He is well qualified to implement these new experiments. However, he has not addressed some fundamental issues in this proposal including-

(1) It is not clear that this approach will benefit the structural genomics initiative. In proteins where spectral overlap is an issue, the introduction of more crosspeaks in the RD-NMR spectrum will pose a real limitation to the usefulness of this approach.

(2) The inclusion of residual dipolar couplings in structure elucidation can be quite useful-however-Dr. Szyperski has not addressed the limitations in this approach.

(3) Sample preparation can be a real issue because of limitations in the solubility of proteins in the bicelle solutions and the narrow temperature range that can be investigated.

Despite the above issues, it is in general a good proposal that would benefit from careful consideration of the potential pitfalls and limitations inherent in this approach. The educational plan is sound and Dr. Szyperski appears to be a very enthusiastic teacher/mentor and deserves recognition for his efforts.

PROPOSAL NO.: 9983995
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Excellent

REVIEW:
MCB-9983995

This CAREER proposal involves the development and application of reduced-dimensionality NMR to protein structure determination with the long-range goal of contributing to structural genomics. A focus will be on reducing the amount of NMR time required for resonance assignment and structure determination of proteins. Dr. Szyperski is one of the original developers of reduced dimensionality methodology when he was a postdoctoral in Wuethrich's lab and thus is uniquely qualified to carry out the proposed research program. A very clever aspect of this research program is to use residual dipolar couplings to aid in sequential assignment of the backbone resonances in a protein. HNCA-type experiments will be modified to include residual dipolar couplings for the alpha protons and carbons and one will be able to help distinguish neighboring residues through differences in dipolar couplings. Thus in the one experiment the residual dipolar couplings will yield powerful additional information for making unambiguous resonance assignments as well as providing data for improving the structure of the molecule. The proposal is well written and the experiments are well outlined. An added strength of the proposal is the collaboration with Dr. Guy Montelione, where Dr. Montelione's AUTOASSIGN program will be use modified to incorporate the reduced dimensionality data. This is a potentially important simplification of the bottleneck in protein structure determination, the resonance assignments.
The PI also does an excellent job of addressing how his research program will be used to enhance aspects of undergraduate and graduate teaching. In addition he proposes that the research program could have broader impact in that he will provide the pulse sequence and assignment protocols to the NMR community and will try to interface with 'NMR parks or consortia' which may be available in the next few years. I believe that the PI has also done an excellent job of trying to address the broader aspects of this proposed research including trying to recruit underrepresented undergraduates into the research program as well as incorporating structural biology into undergraduate

physical chemistry course that he has developed at SUNY Buffalo.
The resources at SUNY Buffalo for this research are outstanding in that Dr. Szyperski has large percentage of the time on both 750 and 600 MHz state-of-the-art spectrometers and all the equipment required for preparation and purification of labeled proteins.

PROPOSAL NO.: 9983995
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Very Good

REVIEW:
The PI has been the leader in the development of 'reduced dimensionality' NMR experiments. In this approach, n chemical shifts are encoded into n$\hat{1}$1 dimensions of a multidimensional NMR spectrum in order to maximize digital resolution while minimizing acquisition time. In the proposed research will have two principle aims: development of RD experiments for measuring dipolar couplings and identification of a minimal set of RD experiments suitable for high throughput structure determination by NMR. Both of these research areas are anticipated by the PI to be important for the use of NMR in structural genomics initiatives.

The PI is well-qualified to perform the proposed experiments and has initiated a valuable collaboration with Prof. G. Montelione (Rutgers) to pursue automated assignments based on RD experiments in conjunction with Prof. Montelione's AUTOASSIGN program.

I am most enthused about the proposed search for a robust minimum set of NMR experiments for structure determination of small protein domains and proteins. Defining such sets will be important for high throughput structure determination in structural genomics initiatives. The collaboration with Prof. Montelione also will enable the selected experiments to be integrated with the strengths of the AUTOASSIGN program. The proposal would be stronger if more discussion was presented of the criteria to be used in recognizing an appropriate set of experiments. How will the PI show that the designed set of experiments is generalizable (to proteins outside the test set) and robust (resistant to common experimental difficulties (poor resolution, disorder, etc.)?

The proposed developments of RD experiments for measuring dipolar coupling constants and of RD experiments incorporating TROSY are less compelling because both of these objectives are straightforward to implement (both the RD and TROSY concepts are generalizable building blocks that can be incorporated into a myriad of

pulse sequences).

The potential widespread application of the proposed methods in structural genomics provides the broad impact envisioned by the NSF's review criterion 2.

The education plan revolves around the teaching of biological multidimensional NMR spectroscopy and structural biology, principally through the new courses CHE 349/350. A strong use of computer and web-based tools is envisioned. The education plan is satisfactorily thought-our and presented, although it is very much in the mainstream of what might be expected of any new faculty member.

PROPOSAL NO.: 9983995
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Very Good

REVIEW:
This application by Dr. Thomas A. Szyperski, proposes to develop approaches to more efficiently and more rapidly elucidate protein structures by using NMR techniques. The proposal addresses both fundamental research and applied research targeting the commercial efficiency of NMR groups focused on structural genomics initiatives.
Dr. Szyperski will focus on implementing a new class of reduced-dimensionality (RD) NMR experiments that allows measurement of residual dipolar couplings in dilute liquid crystalline media. While the determination of residual dipolar couplings from NMR measurements in dilute liquid crystalline media is a novel and potentially useful parameter in structure elucidation, I have some concerns since, for example, various internal motions can play a significant role in the accuracy of residual dipolar couplings. Dr. Szyperski has not addressed this and other issues related to various pitfalls in the research and development of these techniques. Likewise, implementation of the TROSY experiment is commendable; however, will the sample be at least partially deuterated to reduce dipole-dipole interactions? How might this present problems with the expression of various proteins?
Dr. Szyperski proposes to use a set of three proteins (ubiquitin, 8.6 kDa; RNaseA, 15 kDa; and NS-1, 17 kDa). These three proteins are not representative of most proteins that will come from the genome. What about proteins in the 20 kDa to 30 kDa (or higher) where TROSY experiments promise to have the greatest impact in NMR, especially for its use in structural genomics? This has not been addressed. What are the limitations of his eventual protocol for using NMR in structural genomics programs. Even though Dr. Szyperski has the requisite skills in NMR spectroscopy to carry out the proposed work, overall enthusiasm for the proposal has been reduced somewhat by omission of more extensive discussion on experimental pitfalls and alternative approaches.
Dr. Szyperski has developed several very good educational goals, and he has already been active at Colorado in this area. He should be commended for this. In all, this is a very good proposal from an established investigator.

PROPOSAL NO.: 9983995
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: CAREER: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Fair

REVIEW:
The hart of Dr. Szyperski's proposal is to develop a new NMR protocol, with the objective to rapidly determine protein structures for the large number of genes delineated by the sequencing of the genomes.
This appears to be a very worthwhile goal, but this reviewer very strongly doubts whether the proposed research will actually accomplish the desired results. This reviewer is very pleased with the proposed educational plan. Especially the efforts to extend NMR education to all levels of study with appropriate courses, web-based tutorials and hand-on experience is excellent.

The following critique addresses four levels of the research proposal. This reviewer realizes that the critique is harsh, but I hope that this candid assessment will in the long run be helpful to Dr. Szyperski.

1)
At the highest level, Dr. Szyperski needs to address if NMR is the proper method to participate in the process of structural genomics. This premise is just stated in the third line of the project summary and the third paragraph of the background section, but it is not supported by discussion at all. Can NMR really compete with X-ray for rapid turn-over of structures? If so, on what subset of structures? What spectral properties are required?

2)
Depending on the outcome of this question, the next issue should be a discussion of where the bottleneck for NMR structure determination lies.
This reviewer is convinced that it does not lie with the resonance assignment at all. The most time consuming step by far is the NOE identification and iterative structure calculation and refinement. This holds for small proteins as well as large ones. There has even always been a upper size limit for which assignments could still be obtained, but no structure. Therefore, improvement of the efficiency of spectral assignment step

does not really reduce the time necessary for structure determination and is therefore of very limited use to structural genomics.

3)
At a third level, this reviewer differs strongly with Dr. Szyperski's assessment on the efficiency of reduced dimensionality experiments. The experiments are claimed to be more efficient than the current suite of existing experiments. This can be arguably be the case for very small proteins for which no overlap exists in the spectra. However, for the smallest of proteins there is no assignment problem at all, and new methods are not necessary either.
The reduced dimensionality approach places more peaks in the NMR spectrum, which is always a bad idea when spectra get complicated. The next problem is that in experiments such as HNCAHA, there are two Ca frequencies per HN. Thus, the reduced dimensionality experiment will generate 4 peaks for these two peaks for which it is not known how they pair up. It gets worse if there is 2D HN degeneracy. Dr. Szyperski's solution (published by him in the past) to overcome this is to mistune the tau_4 delay in order to obtain axial peaks that contain the HNCA information. But, mistuning will negatively affect the sensitivity of the HNCAHA peaks. The next problem is that in order to obtain high resolution in N-1 dimensions, the indirect time-domain FIDs needs to be collected to high resolution. This leads to low sensitivity of the data. Also, the RD N-1 dimensional experiments cause a splitting of all peaks, which cause a reduction of sensitivity as compared to the N-dimensional experiment. At best, the sensitivity becomes equal if spectra get symmetrized around the center CA positions (which is different for every amino acid). As such, this reviewer is not really happy with the further development of the reduced dimensionality experiments.

4)
At the technical proposal level, I note that the TROSY extension of the RD experiments is technically trivial. Moreover, TROSY is only advantageous for very large proteins, that should not be approached with RD measurements. This limited effort thus seems not so worthwile.

A new approach is to include dipolar couplings in the RD experiments. This is an elegant idea in the context of pulse-sequence design and superficial thinking about automatic assignment stratagems. But, even on the level of spin physics and NMR spectrum appearance, one must wonder if it is desirable to cut the sensitivity of the axial peaks in half by splitting them in two. Of course, one can increase the sensitivity of

the axial peaks by changing the tau_4 tuning, but only at the expense of the sensitivity of the DQ/ZQ peaks. Dipolar doubling of all axial peaks will lead to renewed confusion about the center of the DQ/ZQ HACA doublets the axial peaks were supposed to solve in the first place.

For routine assignment in the genomics context, it cannot be a good idea to try to combine assignment efficiency with dipolar coupling measurement. First, not all proteins dissolve well in bicelle solutions and there are rather stringent temperature and pH requirements. Longevity of samples suffers. Most importantly, 1H linewidths of aligned proteins broaden significantly due to unresolved 1H-1H dipolar interactions, and the sensitivity of all NMR experiments suffers dramatically. In this reviewer's experience it is sometimes even difficult to obtain a decent HSQC (or-IPAP HSQC) on a aligned protein. I strongly doubt if a RD-3D experiment, which already has the very limited sensitivity of a 4D experiment to start with, can be recorded in any reasonable time for an aligned protein. Deuteration to counteract the broadening can of course not be applied for the HNCAHA type experiments. Consequently, I predict that trying to combine measurements for assignment and dipolar couplings will result in that one gets neither.

EXHIBIT 9

PROPOSAL NO.: 0075773
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Very Good

REVIEW:

Szyperski, Thomas

00-75773

Criterion 1: Dr. Szyperski proposes further develop reduced dimensionality (RD) triple resonance NMR methods for use in structural genomics. The research will focus upon development of a protocol that will allow identification of minimal sets of NMR data for structure determination. These advances will be incorporated into automated structure determination schemes to further enhance applicability to a large number of protein targets. The major strengths of the proposal are the expertise of the P.I. in the general area and the importance of the time reductions which potentially could result from the work thereby facilitating determination of a large number of protein structures.

In the first specific aim, existing RD experiments will be implemented and further optimized on Varian spectrometers using Ubiquitin as a test case. This effort appears to be readily feasible and should be accomplished in the early stages of the project. Once implemented, efforts will focus upon implementing RD schemes for measurement of residulat dipolar couplings in liquid crystalline media, definition of a minimal set of spectra which must be acquired for structure determination, evaluation of how automated resonance assignments can be achieved using RD NMR data sets, and finally, development of an "RD-NMR package" for application in other laboratories. All of these represent worthwhile goals and ones that the P.I. has expertise to address. The weakest aspect of the proposal is that much of the work is based upon theoretical predictions and the proposal contains little discussion of potential problems which will have to be overcome.

Criterion 2: The development of methods to facilitate the rapid solution of atomic resolution structures will have a profound impact on the area of structural genomics. Clearly, one of the limitations of structure determination with high-field NMR is the time required for data collection. The studies proposed hold the potential to significantly lessen this time constraint and to lead to the applicability of modern NMR

methods to a larger number of proteins. The applicant is well-qualified to advance this general area an to provide an outstanding training environment of students and post-docs.

Overall Evaluation: Very Good

PROPOSAL NO.: 0075773
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Multiple Rating (Excellent/Very Good)

REVIEW:
This is a very good - excellent proposal by a starting investigator who has made excellent contributions in the field of biomolecular NMR. The proposal is very sound and promises to greatly enhance the utility of solution NMR for structural genomics. For many of the proteins that are likely to be considered by NMR reduced dimensionality NMR is the way to go, offering savings in measuring time. Thus the development of a robust set of experiments is an important step. The applicant has made very important contributions in this area previously and he has the equipment, resources and expertise to continue. I recommend funding.

PROPOSAL NO.: 0075773
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Excellent

REVIEW:
The PI has proposed an orchestrated approach for addressing one of the three major "bottle necks" in the use of high resolution NMR in structure-based genomics. He proposes the development of innovative and novel techniques that would reduce the total instrument time required for obtaining the data needed for sequential resonance assignments and structure determination for small to medium sized proteins. This work dovetails nicely into major research initiatives in the area of functional genomics that are being pursued both here in the US and elsewhere. The PI is superbly qualified and equipped for the proposed studies, with a strong publication record in this area and with ample computer facilities and NMR instrumentation at his disposal.

The proposed studies would make important and relatively novel contributions to the field of structural genomics by developing techniques in two general areas. First, further development of reduced dimensionality (RD) experiments have the potential to dramatically reduce the time required for acquisition and improve the quality of multidimensional NMR experiments for rapid assignment and structure determination. Second, the development of techniques to combine the measurement of residual dipolar couplings with resonance assignment experiments, and to use the variations in residual dipolar couplings to resolve chemical shift degeneracy will further optimize the amount of information extracted from experimental data.

Potential problems seem to be adequately addressed. The most obvious problem in the RD approach is of course the loss of spectral resolution. As the PI points out, the RD triple resonance techniques will benefit from incorporation of TROSY schemes, since the slowly relaxing component selected by TROSY yields a sharper resonance peak and will be of great value in optimizing spectral resolution. This should be adequate for the small to medium sized proteins to which the proposed techniques would be applied.

The PI is taking a collaborative approach, combining the work in his lab with the expertise in the Montelione group on the development of automated assignment tools.

The dissemination of the developed tools is well planned, with versions of the resulting "RD NMR Package" pulse sequences to be made available for both of the major NMR vendors (Varian and Bruker) and data analysis tools that will be compatible with the most widely used data processing packages. Plans for a Web-based infrastructure for teaching and information dissemination should facilitate transfer of the developed technologies to the scientific community.

Finally, this reviewer would like to make a philosophical comment on this and other "high-throughput" efforts connected with the various genomics initiatives. The use of phrases such as "the industrialization of structure determination by NMR" and "protein structure factories" by the PI in the Background section of his proposal brings to mind the Industrial Revolution, and dramatic impact that it had on the human race and its relationship with the natural world. One danger in the use of this terminology is that the highly skilled and creative students trained in the proposed research area will be viewed as technicians, not scientists. Of greater concern, however, is the broader impact of this research on the world at large. How the knowledge gained from genomics research, which will provide a detailed blueprint of all of the components essential to life, will be used is an enormous question. It is the sincere hope of this reviewer that the PI and others in this area of research keep in mind the potential impact of their scientific contributions and keep an open dialog, particularly including their students, on the ethical questions associated with their work. The post-genomic era and the new paradigm shift in scientific research that it brings will have an enormous impact on how we view our world, not unlike the paradigm shift that occurred from the Newtonian world view to that of modern physics. The significance of this shift should not be taken lightly.

PROPOSAL NO.: 0075773
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Very Good

REVIEW:
This is a very good application of a talented scientist who recently joined the faculty of SUNY Buffalo. The proposed work is to utilize and further develop a principle called "reduced dimensionality" (RD) for reducing the measurement time of multidimensional NMR experiments. The PI developed the method while he was in the laboratory of Dr. Wuthrich in Zurich. The idea behind this approach is to record data in a way that both the sum and the difference of the frequencies of a pair of spins is measured along a single indirect dimension so that the information of two dimensions can be read in a single dimension. Thus, one can obtain the information of a 4D experiment essentially in the time one usually spends for a 3D experiment. In more recent implementation, Dr. Szyperski has developed a way to also record the central peak (axial peak), which mey provide additional useful information. The PI claims that this is particularly useful for applications in structural genomics where it is important to acquire spectral information as fast as possible. The proposal also proposes to use this methodology for measuring rsidual dipolar couplings in partially aligned systems. The PI describes convincingly that this information can be obtained from the central peaks in the more recent RD pulse sequences.

This is a technique-oriented proposal. The PI has first described the RD principle seven years ago. Several other groups have applied the technique although it hasn't yet found wide-spread use. However, the benefit of shorter measuring times claimed by the PI is obvious. The reason why the technique hasn't had wide impact may be that it requires additions and subtractions of peak positions to obtain chemical shift data, which complicates analysis of large data sets. This will be less of an issue if the experiments are incorporated in automated assignment routines. Thus, the proposed collaboration with the Montelione laboratory is a very positive aspect of this proposal. Another reason why the RD approach hasn't been used widely is that the primary limitation of NMR structure determination is still the process of making well-behaving protein samples. This may no more be a concern in a structural genomics effort where hopefully many well-behaving proteins will await there structures being solved. In this respect, the

choice of ubiquitin and protein Z and other well-characterized proteins is a little distracting.

Overall, this is a very good application by a new investigator who is a highly talented NMR expert. The technology development proposed has potentially high impact for protein structure determination. The knowledge and research of the PI will have high educational impact on the local structural biology community.

PROPOSAL NO.: 0075773
INSTITUTION: SUNY Buffalo
NSF PROGRAM: MOLECULAR BIOPHYSICS
PRINCIPAL INVESTIGATOR: Szyperski, Thomas A.
TITLE: Reduced Dimensionality NMR Spectroscopy for Structural Genomics
RATING: Very Good

REVIEW:

This proposal focuses on expanding the role of high resolution NMR in a very important, and currently high profile, area of scientific activity, structural genomics. Hypothesis underlying this general area is that the wealth of information coming from sequencing projects can be tapped by solving sufficient numbers of protein structures quickly to provide examples from all fold families (several thousand). X-ray crystallography is clearly the major player in this area, but NMR is important because of its applicability to proteins that do not give diffraction quality crystals. One primary limitation of NMR is that data collection using conventional approaches is slow, requiring on the order of a month of spectrometer time for each protein to be solved. This proposal would expedite the process by identifying a minimal set of NMR experiments, basing these on reduced dimensionality experiments, exploring suitability for automated assignment, and extending schemes to measurement of residual dipolar couplings.

The activity proposed is very useful and it would be carried out under the direction of an investigator with an excellent record. The reduced dimensionality trick, which relies on a proper collection of zero and two quantum coherences, is novel, and one pioneered by the investigator (although there are other examples using simultaneous evolution of two types of chemical shift to reduce dimensions). The proposed interaction with the Montelione group on incorporation of the reduced dimensionality experiments into the AUTOASSIGN program is excellent. And, the extension to collection of residual dipolar data could do much to improve the reliability of assignment and structure determination. There are nevertheless some minor detracting issues. While the reduced dimensionality approach is novel, it is largely implementation that is proposed here - the basic experiments, with the possible exception of those involving residual dipolar couplings, are close relatives of sequences previously published by the investigator. And, while the estimated reduction in acquisition time (a week) is certainly significant, one wonders whether more radical departures from existing protocols might be more productive than refinement of existing ones.

The training opportunities in the proposed work are substantial in the area of NMR spectroscopy and the investigator would seem to be an excellent mentor of graduate students in this area. The area of application is very exciting, and should appeal to prospective graduate students. It is a little disappointing that educational plans are not broader, taking advantage of the obvious link to genetics and protein function.

Summary: This is a very good proposal. It addresses a very timely, important, problem. The investigator is highly skilled and qualified to conduct the proposed research. While the work is based on novel experiments previously publsihed by the investigator, the fact that the focus of the current proposal is more on implementation than new innovation moderates the overall level of enthusiasm.

EXHIBIT 10

NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230

DIVISION OF MOLECULAR AND CELLULAR BIOSCIENCES

Dr. Thomas A. Szyperski
Department of Chemistry
SUNY at Buffalo - Amherst Campus
816 Natural Sciences Complex
Buffalo, NY 14260

Ref: MCB-0075773

Dear Dr. Szyperski:

We are very pleased that the National Science Foundation will support your proposal entitled: "Reduced Dimensionality NMR Spectroscopy for Structural". In addition to extending our congratulations, there are several purposes for this letter.

First, we want to make you aware of several special opportunities for supplementing your NSF grant. A one page description of opportunities for supplemental funding is included with this letter. Please note that all supplements are contingent upon availability of funds. The Federal fiscal year is October 1-September 30. Requests for a particular fiscal year's funds should be submitted directly to the Program by April 1. It may not be possible to honor requests received later in the fiscal year.

Second, we strongly encourage you to keep the Program informed of any particularly significant and interesting results as your work progresses. For example, please let us know after peer review but prior to publication of any upcoming *Science*, *Nature*, or other publication which represents an important research accomplishment. This will enable us to do a better job of communicating the excitement of modern biology and its contributions to the general public and in justifying our annual budget request Please remember that it is important that you acknowledge NSF support in talks and poster presentations as well as on publications resulting from your award.

Finally, please remember that to process future yearly increments of your award, your annual progress report must be received by the program at least 60 days before the anniversary date of this award OR by April 1, whichever is earlier. The format for the summary of scientific progress is available on the project reporting system in FastLane. You will be sent further information on this at a later time.

Copies of the reviews of your proposal and a *Panel Summary,* outlining the salient points raised in the panel discussion of your proposal, are enclosed. These are for your personal use and are not made available by the Foundation to anyone else. They may be helpful to you in the conduct of your research or in preparing future proposals.

Please feel free to contact me if you have any questions about these or other matters concerning your NSF grant. We wish you every success.

Sincerely,

Thomas E. Smith, Ph.D.
Molecular Biochemistry
Molecular & Cellular Biosciences
Phone: (703) 306-1443/Fax: (703) 306-0355
E-mail: tesmith@nsf.gov

EXHIBIT 11

1st submission

Journal of the
American Chemical Society

Department of Chemistry
University of Arizona
1306 E. University Boulevard
Tucson, AZ 85721-0041
Phone: (520) 626-9309
FAX: (520) 626-9300
E-Mail: jacs@u.arizona.edu

PUBLISHED BY
THE AMERICAN CHEMICAL SOCIETY

Dr. F. Ann Walker, *Associate Editor*

July 17, 2001

By email to <szypersk@nsm.buffalo.edu> and by regular mail

Dr. Thomas Szyperski
Chemistry and Biochemistry
State University of New York at Buffalo
Department of Chemistry
816 Natural Sciences Complex
Buffalo, NY 14260

Ms No.: ja016178i Security Key: J946 (both case sensitive)
Title: "Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis"

Dear Dr. Szyperski:

Your manuscript was sent to three reviewers, two of whose comments are enclosed. Both of the reviewers feel that the work is appropriate for publication in the Journal, but only after a number of major points are addressed. I will be willing, then, to reconsider this paper, revised in light of the reviewers' comments. In your cover letter, discuss the changes made to address the reviewers' comments.

Please submit your revised manuscript using the JACS website submission process beginning at <http://pubs.acs.org/cgi-bin/submission_gen/index.pl?Journal=jacsat>, choosing either Option 1 (author supplied PDF files) or Option 2 (system generated PDF files). However, rather than clicking on "Proceed" key in your permanent manuscript number and the security key shown above.

Thank you for submitting this manuscript to *J. Am. Chem. Soc.*

Sincerely yours,

F. Ann Walker

F. Ann Walker
Associate Editor

Reviewer 1

Manuscript number: ja016178i
Manuscript title: Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis
Corresponding author: Thomas Szyperski

Recommendation: Publish after major revisions.

(1) Is the manuscript likely to be of interest to the broad readership? no
(2) Are the conclusions adequately supported by the data presented? no
(3) Are the literature references appropriate and correct? yes
(4) Does the nomenclature used conform with accepted practice? yes
(5) Are hazardous procedures clearly defined as such? yes

General Comments:

SUMMARY

This paper is largely a statement of advocacy rather than a critical scientific evaluation and account. While there is certainly merit in reduced dimensionality experiments for accelerating resonance assignments, the case study of a 4.5 ns correlation time polypeptide, the projected enhancements with cryoprobes, single transient spectra, etc. are not only misleading for the general scientific readership but scientifically indefensible with the data shown. The paper reads more like an advertising campaign, or a grant proposal on what *will be* done rather than what *is* reported. For example the statement (pp 19) "will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less" is a projection of idealized measurement time, a zero-time processing and peak-picking allocation, on an extremely idealized protein. It does not convey a realistic picture.

Thus, the paper would only be acceptable after major revision. This includes shortening the paper to remove much of the redundancy in statements and in clearly delineating the **results** from **predictions**.

Some specific points requiring attention (though there are more):

ABSTRACT

resonance assignments are required for all spectroscopic interpretation, not just structure and SAR by NMR.

"practical utility" needs more than an 8.5kDa domain with a 4.5ns correlation time. How would these spectra fare with a 15ns correlation time protein? What subset of spectra could be acquired? Would they still be sampling limited? How robust is analysis to missing data?

I fail to see how "protein secondary structure will support protein fold prediction". Secondary structure is akin to identifying the logs used to make a log cabin - it tells us nothing about the 3D configuration of the cabin - that is the fold.

Introduction

There needs to be a discussion of alternative methods of dealing with "sampling limited" NMR schemes. Sampling limitations hinge on the requirement of the Fourier transform to have uniformly incremented sampling. Other mathematical methods such as Filter Diagonalization and Maximum Entropy Method processing can circumvent this limitation. In addition spectral aliasing or folding can increase the effective dwell time and sample a given dimension with the same resolution with considerably fewer points. In fact, the RD method appears to *decrease* the dwell time to account for the two frequencies being measured, and this needs discussion. Table 2 shows (for example) the HabCab(CO)NHN with a tmax of 6.3ms and 95* points which translates into a C/H spectral width of > 15,000Hz. In a non-RD experiment a typical 13C spectral width would be 2000-4000Hz depending on the extent of aliasing, and possibly 1000-2000Hz for 1H. At issue is whether the total number of increments in an nD spectrum vs an (n-1)D RD experiment for equivalent digital resolution - it most likely will be but both sides of the argument need to be presented. Since the sensitivity for an nD spectrum is identical to an (n-1)D RD spectrum alternative sampling schema must be addressed. And it should be addressed in the context of increased overlap in the RD spectra that would not be present in a higher-D spectrum.

I am a bit surprised that what normally would be considered spectral artefacts of a transfer process (axial coherences from incomplete INEPT or Boltzmann differences) are extolled as a virtue (pp4).

Materials and Methods

Several of the proposed experiments will suffer from signal loss in more typical proteins, for example, 1C, 1J, 1K and will not be appropriate for larger proteins. Even with 4.5ns correlation time signals are already missing from the HC-(C-TOCSY-CO)NHN. Additionally, while the first 6 experiments are HN detected experiments in H2O are required, but in experiments H and I where Ha are detected in the presence of 100M protons in water what is done to suppress that overwhelming signal?
Were the experimental results shown really done with a single transient per FID, or is this simply used to calculate the minimum *theoretical* acquisition time? Why do the HN detected experiments not have the water magnetization retained along the z-axis?

Results and Discussion

Why are the traces in spectra in Figure 4B, 4C,4E,4F, 5A,5B,5C out of phase- particularly why are HACA affected differently from HBCB etc.? What is the origin of the artifact at 73ppm in 4F? and why is Cb missing from 4F yet present in 4C? Artifacts and phase errors will limit the utility of spectra, particularly in peak discrimination and in measurement precision which are central to reliable assignments.

Tale S2 has the detection yield ratio inverted and t_{mean} rather than t_{meas}. While experimental statistics are helpful I think the "average" S/N is not a good measure. More preferable would be the "spread" which points out if some residues are site-specifically weaker. Total assignment relies on detecting the lowest sensitivity cross-peaks, and in the case of exchange broadened spectra sensitivity enhanced spectra may actually get worse for these residues.

The statement on pp19 "will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less" gives no regard to processing the spectra, and RELIABLY peak-picking (discerning signal not only from thermal noise but also artifact noise) the data before sorting in AutoAssign.

pp20 sweep width should be replaced with spectral width.

Reviewer 2

Manuscript number: ja016178i
Manuscript title: Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein
Resonance Assignment: Implementation and Automated Analysis
Corresponding author: Thomas Szyperski

Recommendation: Publish after major revisions.

(1) Is the manuscript likely to be of interest to the broad readership? Not applicable or unknown
(2) Are the conclusions adequately supported by the data presented? Not applicable or unknown
(3) Are the literature references appropriate and correct? Not applicable or unknown
(4) Does the nomenclature used conform with accepted practice? Not applicable or unknown
(5) Are hazardous procedures clearly defined as such? Not applicable or unknown

General Comments:

The manuscript of Szyperski et al. presents a suite of reduced dimensionality triple resonance experiments for the rapid assignment of the backbone and side chain resonances of small to medium-sized proteins. A major part of the discussion deals with a critical comparison of the sensitivity of the individual experiments and the optimisation of measuring time taking into account the requirements of high spectral resolution and sufficient signal to noise. In addition, analysis of these NMR data has been implemented in the program package AutoAssign. Although I believe that the manuscript contains some interesting ideas, I am sceptical whether it merits (in the present form) publication in the Journal of the American Chemical Society.

The authors present eight unpublished pulse sequences for triple resonance experiments using the so called reduced dimensionality approach. These experiments are well presented with a comprehensive description of the different pulse sequences in the Supporting Information. This part of the manuscript is certainly of interest to the biomolecular NMR community and merits publication in a specialised NMR journal.

The major aim of the manuscript, however, is to present a new general strategy for speeding up the NMR assignment step of proteins. Reduced dimensionality experiments provide connectivities between four different nuclei and the high spectral resolution required for an automated data analysis. In addition, the authors present a comparison of the sensitivity of eleven NMR experiments recorded on a small protein of 63 residues at a magnetic field strength of 14.1T, and discuss theses results in terms of "minimal required data set" and "minimal required experimental time". Although I believe that NMR assignments will greatly benefit from the use of reduced dimensionality experiments, I have some major concerns about the conclusions drawn from the experimental results. The conclusions about the relative sensitivity of the NMR experiments are certainly valid for other proteins with similar molecular weight (well below 100 residues) and studied under

similar experimental conditions (temperature, magnetic field strength), but they will (as the authors agree on page 13) completely change for larger molecules (higher tumbling correlation times). Along the same lines, the chosen set of eleven triple resonance experiments may (or may not) be the best choice for assignment of a small protein, but other experiments will certainly yield much better results in the case of larger perdeuterated or randomly fractional deuterated proteins. Finally, a combination of reduced dimensionality experiments (for the most sensitive ones) and standard 3D triple resonance experiments (for the less sensitive ones) seems to me a better choice than the one presented in the manuscript. Thus the major conclusions of the manuscript are rather subjective and far from being general. By the way, the statement that complete NMR assignment of medium sized molecules will be possible "within a day or less" (page 19) is a very optimistic statement. It could even have a negative impact on most of the biomolecular NMR laboratories, where this step still requires a couple of weeks up to several months. This sentence should therefore be dropped from the manuscript unless the authors prove that they are really capable of what they are claiming.

In conclusion, for the manuscript to be acceptable for publication in JACS, I suggest that the authors add at least one additional experimental study on a second test molecule (in the range of about 150 residues). This will help the work to be of more general interest to the readership of JACS. As recording of the NMR data sets and assignment can be accomplished in a couple of days by the authors, this requirement should not significantly delay publication of this work.

Further points concerning the presentation of the work:

1. the identification of the secondary structural elements is presented as a second additional information which can be rapidly obtained by the proposed strategy. This is misleading as the determination of secondary structure on the basis of ^{13}C chemical shifts is just a consequence of the assignment (and therefore trivial).
2. the last part of the Conclusions section (page 21) reads more like a grant application than a scientific presentation and should be dropped from the manuscript.
3. T. Szyperski is complete in citing his own research papers on reduced dimensionality triple resonance experiments, but rather selective in citing the work of other groups. References to (at least) the following two papers should be added: (1) Simorre et al (1994) J. Biomol NMR 4, 325; (2) Pang et al (1998) J. Biomol NMR 11, 185.

Minor points (typos):

- page 6: the sample conditions were probably 90% H2O and 10% D2O (and not the inverse)

EXHIBIT 12

Dr. T. Szyperski
Associate Professor of Chemistry and Biochemistry
Department of Chemistry, 816 Natural Sciences Complex
University at Buffalo, The State University of New York, Buffalo, NY, 14260
Phone: 716-645-6800 ext2245, Fax: 716-645-7338
E-mail: szypersk@acsu.buffalo.edu

Dr. F. Ann Walker
Associate Editor, Journal of the American Chemical Society
The University of Arizona
Department of Chemistry
1306 E. University Blvd
Tucson,
Arizona 85721-0041

ja016178i

8/8/2001

Dear Prof. Walker,

thank you very much for your letter of July 17, 2001. We are delighted to hear about the positive response with respect to our publication.

Please find enclosed the revised version of our manuscript entitled "Reduced-dimensionality NMR spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis", by T. Szyperski, D. Yeh, D. Sukumaran, H. Moseley and G. Montelione.

To meet with the concerns of the reviewers, we have extensively revised the Ms. to avoid claims beyond what is actually shown in the paper. Moreover, we have (i) substantially shortened the Ms. to eliminate redundancy, (ii) largely rewritten the section "prospects for larger proteins", (iii) extended the supplementary material, and (iv) clearly delineated the results from the predictions.

I have attached a detailed list of changes that we made to address the reviewers' comments.

We hope that the revised version is acceptable for publication in the *Journal of the American Chemical Society.*

Thank you very much for your efforts.

Yours sincerely,

Thomas Szyperski

Reviewer 1:

I....with respect to the **SUMMARY**:

1....."....The paper reads like an advertising campaign, or a grant proposal.....
We have extensively revised the "Results and Discussion" and "Conclusions" sections (pages 11-19). to avoid claims beyond what is actually shown in the paper. We have also substantially shortened the Ms.

2....."For example, the statement "will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less"....."

We have deleted the sentence "Hence, we conclude that the joint employment of RD NMR spectroscopy, automated backbone resonance assignment and cryogenic probes will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less." on former page 19.

II.....with respect to the **ABSTRACT**:

1. "resonance assignments are required for all spectroscopic interpretation, not just structure and SAR by NMR"

We wrote..."is presented for high-throughput resonance assignment of proteins, which is required for.....". High-throughput resonance assignment is *not* required for all spectroscopic interpretation.

To further focus our statement, we have shortened the first sentence of the abstract, now solely referring to high-throughput assignment for structural genomics.

2...."practical utility" needs more than an 8.5 kDa domain...
Studies of proteins tumbling with a correlation time of 4.5 ns are of great practical utility. Even if the method *would* be limited to 5-7 ns, one needs to recognize that currently most biological NMR studies are pursued with systems tumbling with less than 10 ns. (see also next point).

3...."How would these spectra fare with a 15 ns..."
This is discussed in the separate section "Prospects for larger proteins" and in the "Conclusions" (pages 11-19). To further clarify this, we have rewritten the section the section "Prospects for larger proteins" (page 16, 2nd para,; page 17, 1st para). Note, that the prospects of 3D RD NMR are basically those of 4D TR NMR. [Lewis Kay has published 4D TROSY TR spectra for a system tumbling with 46 ns (Yang and Kay, *J. Am. Chem. Soc.* **121**, 2571-2575)]. This view is backed by the outstanding short measurement times that were employed for the present study.

4..."What subset of spectra could be acquired?"
The proposed subsets for larger (as well as smaller) proteins are outlined in detail on pages 13-15 in the sections "A 'standard set' of RD NMR experiments", "A 'minimal set' of RD NMR experiments" and "Prospects for larger proteins". In view of the reviewers request to shorten the Ms, we preferred not to expand the abstract.

5..."Would they still be sampling limited?"
This certainly depends on (i) protein concentration, (ii) spectrometer sensitivity and (iii) field strength (at ultra high field the spectral widths increases). We have addressed that issue in the "Prospects for larger proteins" and in the "Conclusions" (pages 16-19). In view of the reviewers request to shorten the Ms, we preferred not to expand the abstract.

6..."I fail to see how "protein secondary structure will support protein fold prediction"...."

Fold prediction based on known secondary structure is far more reliable than "ab initio" approaches relying on the polypeptide sequence only. To clarify this point, we have introduced a new reference 53 (Ortiz et al.,) in the "Conclusions" section of the manuscript (page 19, first 2 lines).

III....with respect to the **INTRODUCTION**:

1...."There needs to be a discussion of alternative methods of dealing with "sampling limited" NMR schemes. Sampling limitations hinge on the requirement of the Fourier transform to have uniformly incremented sampling. Other mathematical methods such as Filter Diagonalization and Maximum Entropy Method processing can circumvent this limitation."

We have alluded to those methods in the "Conclusions" section (page 19, 2nd para, lines 13-17), and the corresponding publications are quoted, i.e., we suppose that no change of the Ms. is required. Importantly, the practical utility of, for example, "Filter Diagonalization" has not yet been demonstrated, and remains a perspective. In contrast, the utility of RD NMR spectroscopy has been demonstrated for several systems.

2...."In addition spectral aliasing or folding can increase the effective dwell time and sample a given dimension with the same resolution with considerably fewer points. In fact, the RD method appears to *decrease* the dwell time to account for the two frequencies being measured, and this needs discussion. Table 2 shows (for example) the HabCab(CO)NHN with a tmax of 6.3ms and 95* points which translates into a C/H spectral width of > 15,000Hz. In a non-RD experiment a typical 13C spectral width would be 2000-4000Hz depending on the extent of aliasing, and possibly 1000-2000Hz for 1H. At issue is whether the total number of increments in an nD spectrum vs an (n-1)D RD experiment for equivalent digital resolution - it most likely will be but both sides of the argument need to be presented. Since the sensitivity for an nD spectrum is identical to an (n-1)D RD spectrum alternative sampling schema must be addressed. And it should be addressed in the context of increased overlap in the RD spectra that would not be present in a higher-D spectrum."

In conventional CBCA(CO)NH, a spectral width on 10,000 Hz is usually chosen along w1(13c) at 600 MHz, and this dimension is, quite generally, not aliased. A 1H spectral width of 2,000 Hz is reasonable in a HBHA(CO)NH experiment. Since the 1H carrier is set to the edge of the spectral range in the projected dimension, we thus chose sw(13c) + 2*sw(1h) ~ 15,000 Hz. It takes 190 FIDs (95 complex points) in the RD experiment to sample w1(13c/1h). Assuming sweep widths as indicated for the 4D HBCBHACA(CO)NHN experiment, one would need 64 complex points in w2(13c) and 13 complex points in w1(1h), i.e., 2*64*2*13 = 3,328 FIDs to sample both dimentions. Hence, the gain in minimal measurement time, in spite of an increased sweep width in the projected dimension in 3,328/190 = 17. This cannot be compensated for by aliasing.

In general, the reviewer is referring to the gain in minimal measurement time upon projection. As a rule of thumb the gain is about an order of magnitude. The formula to calculate the gain are given in the Appendix of Ref 20 (Szyperski, T.; Banecki, B.; Braun, D.; Glaser, R. W. *J. Biomol. NMR* **1998**, *11*, 387–405.). The large gain is due to the fact that the sampling of two dimensions with i and j complex points, is proportional to i times j.

As an illustration considering aliasing in the conventional experiment, one may compare a 3D RD HCCH and a 4D HCCH experiment:

3D HCCH-COSY	4D HCCH-COSY
For the combined dimensions:	
$t_{3,max}(^{13}C/^{1}H)$ = 6.6 ms	$t_{3,max}(^{13}C) = t_{4,max}(^{1}H)$ = 6.6 ms
SW2($^{13}C/^{1}H$) = 12,000 Hz	SW3(^{13}C) = 3,000 Hz SW4(^{1}H) = 3,000 Hz
80 complex points = 160 FIDs	20 * 20 complex points = 1,600 FIDs
For the entire data set:	
20 * 80 complex points= 6,400 FIDs (1.8 h)	20 * 20 * 20 complex points = 64,000 FIDs (18 h)

=> Reduction in minimal measurement time @ 600MHz: factor ~10

Here, aliasing was assumed in w2(13c) and w3(13c) of the 4D experiment, while no aliasing was assumed for the RD experiment.

To further clarify this point, we have modified the fourth paragraph of the introduction (page 4, 2nd para, lines 6 and 7) accordingly.

Since all these points were previously published and described in great detail, and in view of the reviewers request to shorten the Ms., we have not further extended the Ms.

3....."I am a bit surprised that what normally would be considered spectral artefacts of a transfer process (axial coherences from incomplete INEPT or Boltzman differences) are extolled as a virtue (pp4)."

This "virtue" was published in a communication to the *Journal of the American Chemical Society* in 1996 (Ref. 13, Szyperski, T.; Braun, D.; Banecki, B.; Wüthrich, K. *J. Am. Chem. Soc.* **1996**, *118*, 8146–8147.), which was entitled:
"*Useful Information from Axial Peak Magnetization in Projected NMR Experiments*"

IV..... with respect to **MATERIALS AND METHODS**

1....."Several of the proposed experiments will suffer from signal loss in more typical proteins, for example, 1C, 1J, 1K and will not be appropriate for larger proteins."

With respect to 1C, we explicitly advertise this experiment for smaller protein (page 15, last 3 lines), and, in order to avoid additional redundancy, did not modify the Ms.
With respect to 1J, we have extended the Ms. on page 16 (2nd para, lines 5-8) pointing at the combined use with NOESY in the paragraph entitled "Prospects for larger proteins"..
With respect to 1K: this is a very sensitive tailored for large proteins (in fact, its sensitivity is higher than HCCH-COSY: see Ref 38, Zerbe, O.; Szyperski, T.; Ottiger, M.; Wüthrich, K. *J. Biomol. NMR* **1996**, 7, 99–106.). See also Fig. 3.

2....."Even with 4.5 ns correlation time signals are already missing from the HC-(C-TOCSY-CO)NHN."

We discuss this on page 11 (last 4 lines) and page 12 (first 5 lines). Please note, that completeness of observation is not expected even for small proteins because the TOCSY spin modes generate zero net transfer between certain carbons at given mixing times (see also below), and because relay to the most remote carbons in the long side chains would require an unrealistically long mixing time. In view of the request to shorten the Ms., we have not further extended this discussion.

3...."Additionally, while the first 6 experiments are HN detected experiments in H2O are required, but in experiments H and I where Ha are detected in the presence of 100M protons in water what is done to suppress that overwhelming signal?"

For *non* HN detected experiments, the water signal is suppressed by coherence rejection using spin-lock purge pulses and pulsed filed z-gradients. For HN detected experiments water suppression is primarily based on the coherence selection method of Kay. No presaturation of the water line was applied, except for the experiments detecting magnetization on the aromatic protons. To further clarify this issue, we have extended the corresponding legends of the pulse schemes of the Supplementary accordingly (see also 5.).

4....."Were the experimental results shown really done with a single transient per FID, or is this simply used to calculate the minimum *theoretical* acquisition time?"

We wrote "If the standard (or minimal) set of experiments *would have been* recorded with a single transient per increment,..." (page 17, last three lines). Hence, we thought it is obvious that the data were not acquired with a single transient per FID (We have, however, recorded single transient spectra for several proteins of the Northeast Structural Genomics Consortium's NMR structure pipeline). To further clarify this point, we have modified the Ms. on page 17 (2nd para), we have taken the right-most column out of Table 2 (to also contribute to clearly delineate results and predictions), and we have added a figure in the supplementary material (Fig. S20) that shows the excellent quality of single transient detection that is nowadays possible.

5....."Why do the HN detected experiments not have the water magnetization retained along the z-axis?"

The experiments were conducted at pH = 6.5, where signal attenuation for HN detected experiments arising from exchange with saturated water magnetization is quite small (which is evidenced by the excellent signal-to-noise ratios measured in the HN detected experiments; Table S2). To deal with this point, we have extended the methods part accordingly (page 7, 2nd para, lines 9-11).

V.....with respect to "**RESULTS AND DISCUSSION**"

1....."Why are the traces in spectra in Figure 4B, 4C, 4E, 4F, 5A, 5B, 5C out of phase - "

We do not see that 4C, 4E, 4F, 5A, 5B and 5C are (within the signal-to-noise) out of phase. The slight phase distortion in 4B did not impede manual or automated data analysis.
Importantly, the symmetry properties of RD TR NMR spectra (peak pairs need to be identified) greatly facilitate to distinguish between peaks at the noise level and artefacts. We have introduced this point out on page 10, lines 7-10.

2....."particularly why are HACA affected differently from HBCB etc.?"

Because the r.f. pulse schemes are different (carrier positions, pulse lengths), yielding different imperfections from, for example, off-resonance effects (the spectral artefacts in conventional CBCA(CO)NHN and (HA)CA(CO)NHN are likewise different).

3....."What is the origin of the artifact at 73ppm in 4F?"

The putative artefact is, if at all, within the noise level.

4....." and why is Cb missing from 4F yet present in 4C?"

Because of the particular mixing time chosen (the signals are strong at the shorter mixing time of 14 ms). We have alluded to this point in our Ms. on page 12, lines 1-4 (referring also to Fig 4). To further clarify this point, we have extended the legend of Fig. 4 accordingly (see also IV.2).

5..."Artifacts and phase errors will limit the utility of spectra, particularly in peak discrimination and in measurement precision which are central to reliable assignments."

Neither manual nor automated analysis was impeded by any of the small artefacts in the spectra (which are likewise present in conventional TR spectra). See also the high-quality of the spectra shown in Figs. 7, 8, 9, S12 and S13. Importantly, the symmetry properties of RD TR NMR spectra (peak pairs need to be identified) greatly facilitate to distinguish between peaks at the noise level and artefacts. We have introduced this point out on page 10, lines 7-10.

6....."Table S2 has the detection yield ratio inverted and t_{mean} rather than t_{meas}."

We have inverted the ratios, and we have corrected the typo in header of Table S2.

7....."While experimental statistics are helpful I think the "average" S/N is not a good measure. More preferable would be the "spread" which points out if some residues are site-specifically weaker. Total assignment relies on detecting the lowest sensitivity cross-peaks, and in the case of exchange broadened spectra sensitivity enhanced spectra may actually get worse for these residues."

We agree with the reviewer, and have included 22 signal-to-noise distributions (Figures S14 to S19 in the Supplementary Material (3D HNNCAHA, HabCab(CO)NH, HabCabCOHA, HNN<CO,CA>, HCCH-COSY) covering backbone and aliphatic side chains. The detection of the (relatively fewer) aromatic signals is comprehensively documented in Fig. 9.

8....."The statement on pp19 "will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less" gives no regard to processing the spectra, and RELIABLY peak-picking (discerning signal not only from thermal noise but also artifact noise) the data before sorting in AutoAssign."

We have deleted the sentence "Hence, we conclude that the joint employment of RD NMR spectroscopy, automated backbone resonance assignment and cryogenic probes will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less." on former page 19.

9....."pp20 sweep width should be replaced with spectral width."

We have introduced "spectral width" on former page 20.

Reviewer 2:

I. General comments

1....."The conclusions about the relative sensitivity.....and studied under similar experimental conditions (temperature, magnetic field strength), but they will (as the authors agree on page 13) completely change for larger molecules (higher tumbling correlation times)."

This is discussed in the separate section "Prospects for larger proteins" and in the "Conclusions" (pages 11-16). To clarify this issue, we have largely rewritten the section "Prospects for larger proteins". See also I.1, II.2, II.3 of reviewer 1.

2....."Along the same lines, the chosen set of eleven triple resonance experiments may (or may not) be the best choice for assignment of a small protein, but other experiments will certainly yield much better results in the case of larger perdeuterated or randomly fractional deuterated proteins."

We are describing a new strategy to obtain complete resonance assignments for proteins using RD NMR spectroscopy, and we describe in detail the impact of protein deuteration (page 17, entire 1st para). Since protein deuteration affects the performance of RD NMR TR experiments in the same fashion as conventional TR experiments, the impact of deuteration on RD NMR in known and does not require a separate investigation. Please note also, that, due to the request of reviewer 1 to shorten the Ms., we cannot extent the Ms. incorporating a detailed (review-like) comparison with the larger number of TR assignment strategies that are currently used.

3...."Finally, a combination of reduced dimensionality (for the less sensitive ones) seems to me a better choice than the one presented in the manuscript."

This point is already made in the paper, i.e., we agree that a combination of RD with conventional TR experiments is a viable choice. In particular, we explicitly propose to use conventional HNNCACB in conjunction with the suite of RD schemes (e.g., Figure 2).

4....."Thus the major conclusions of the manuscript are rather subjective and far from being general."

We have extensively revised the Ms. to avoid claims beyond what is actually shown in the paper. We believe that the major conclusions presented in the revised paper are general to the very best of our knowledge.

5....."By the way, the statement that complete NMR assignment of medium sized molecules will be possible "within a day or less" (page 19) is a very optimistic statement. It could even have a negative impact on most of the biomolecular NMR laboratories, where this step still requires a couple of weeks up to several months. This sentence should therefore be dropped..."

....We have deleted the sentence "Hence, we conclude that the joint employment of RD NMR spectroscopy, automated backbone resonance assignment and cryogenic probes will allow one to determine a protein's backbone resonance assignments and secondary structure within a day or less." on former page 19.

6....." In conclusion, for the manuscript to be acceptable for publication in JACS, I suggest that the authors add at least one additional experimental study on a second test molecule (in the range of about 150 residues)."

The suite of RD NMR experiments presented here has been employed for proteins in the NMR structure pipeline of the Northeast Structural Genomics Consortium with the following size distribution: 8, 12, 12.5, 14, 15, 22 kDa (see also new Fig. S20 In all cases, the total measurement time amounted to about a week or less for the entire standard set of experiments. To address this concern, we have introduced this information in the section "prospects for larger proteins" (page 16, 2nd para, lines 7-12). The results of these RD NMR measurements within the consortium are, in part, currently written up, and evidently involve a larger number of additional co-authors. Hence, these data cannot just be included in this manuscript. Please also note that
(i) for the assignment of the 21 kDa protein FimC (180 residues), the Ha/bCa/b(CO)NHN experiment was recorded back in 1997 on a 600 MHz spectrometer [J Biomol NMR, 1998, Feb;11(2):229-30; "Sequence-specific 1H, 15N, and 13C assignments of the periplasmic chaperone FimC from Escherichia coli"; Pellecchia M, Guntert P, Glockshuber R, Wuthrich K]; this paper as been introduced as the new reference 50 (page 16, 2nd para, first 2 lines),
and (ii) the prospects of 3D RD NMR for large proteins are basically those of 4D TR NMR (see also: II.3 of reviewer 1).

The reviewer's concerns are now addressed in the largely rewritten paragraph ("prospects for larger proteins").

II....."Further points concerning the presentation of the work:"

1. ....."the identification of the secondary structural elements is presented as a second additional information which can be rapidly obtained by the proposed strategy. This is misleading as the determination of secondary structure on the basis of ^{13}C chemical shifts is just a consequence of the assignment (and therefore trivial)."

Although it is straightforward to derive a protein's secondary structure for the $^{13}C^{\alpha}$ chemical shifts, this is key for HTP and structural genomics. We think, that it is not at all misleading, but simply correct to point out that the resonance assignment itself provides the secondary structure. However, to clarify this point, we have modified the Ms. accordingly (page 19, line 3).

2. ....."the last part of the Conclusions section (page 21) reads more like a grant application than a scientific presentation and should be dropped from the manuscript."

We have deleted the last paragraph of the "Conclusions" section.

3....."T. Szyperski is complete in citing his own research papers on reduced dimensionality triple resonance experiments, but rather selective in citing the work of other groups. References to (at least) the following two papers should be added: (1) Simorre et al (1994) J. Biomol NMR 4, 325; (2) Pang et al (1998) J. Biomol NMR 11, 185."

We have introduced Ref (1) as the new Ref. 12b (joining two other references [12a and 14] from the Marion group that were already quoted). We could, however, not include the proposed Ref (2), simply because this paper does not describe a reduced-dimensionality NMR experiment.

4....."Minor points (typos)"
- page 6: the sample conditions were probably 90% H2O and 10% D2O (and not the inverse)

We have corrected this typo.

EXHIBIT 13

Journal of the
American Chemical Society

PUBLISHED BY
THE AMERICAN CHEMICAL SOCIETY

Department of Chemistry
University of Arizona
1306 E. University Blvd.
Tucson, Arizona 85721-0041
Phone: (520) 626-9309
Fax: (520) 626-9300
Internet: jacs@u.arizona.edu

Dr. F. Ann Walker

August 13, 2001

By mail and fax 716 645 7338

Dr. Thomas Szyperski
Chemistry and Biochemistry
State University of New York at Buffalo
Department of Chemistry
816 Natural Sciences Complex
Buffalo, NY 14260

Ms No.: JA016178I Web Article
Title: "Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis"

Dear Dr. Szyperski:

Your revised manuscript was sent back to Reviewer 1, whose comments are enclosed. Unfortunately, this reviewer feels that the revised manuscript is not appropriate for publication in the Journal, but rather should appear in a more specialized journal. In this situation I have no alternative but to reject the manuscript, and suggest that you follow the suggestions of the reviewer in preparing a revised manuscript for submission elsewhere.

Thank you for submitting this manuscript to *J. Am. Chem. Soc.*

Sincerely yours,

F. Ann Walker
Associate Editor

Reviewer 1

Manuscript number: ja016178i
Manuscript title: Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein
Resonance Assignment: Implementation and Automated Analysis
Corresponding author: Thomas Szyperski

Recommendation: Publish elsewhere
Journal suggested: J. Bio. NMR or J. Magn. Reson.

(1) Is the manuscript likely to be of interest to the broad readership? no
(2) Are the conclusions adequately supported by the data presented? no
(3) Are the literature references appropriate and correct? yes
(4) Does the nomenclature used conform with accepted practice? yes
(5) Are hazardous procedures clearly defined as such? yes

General Comments:

I regret that the changes proposed in the manuscript in the way of explanations and answers lead me to the conclusion that this paper is *not* suitable for publication in JACS but rather should be submitted to a specialist journal where it can be reviewed by other experts. Like Reviewer 2 I still feel that the paper makes some rather optimistic statements and would mislead the JACS audience-at-large about efficiency and tractability of NMR assignments. Although the authors have shortened the manuscript it is still very extensive and seems to offer too much detail where it is probably not needed, yet insufficient where it is needed. A simple example is the stick figure representation of Figure 1A showing H α/β C α/β (CO)NHN where I am sure most chemists would wonder what happened to the other (geminal) Hβ.

The questions still raised are matters of degree. For example in the response cover letter, "Studies of proteins tumbling with a correlation time of 4.5 ns are of great practical utility. Even if the method would be limited to 5-7 ns, one needs to recognize that currently most biological NMR studies are pursued with systems tumbling with less than 10 ns (see next point)" let me just add the following line of arithmetic:

$$\exp -(10\text{ns}/4{,}5\text{ns}) = 0.11.$$

That is, a 10 nsec correlation time will reduce sensitivity by approximately an order of magnitude over the example given. The "(see next point)" is a reference to the *most* sensitive of triple resonance experiments and applied only to a backbone rather than the least sensitive, with an emphasis on total assignment. The distinction between "small" and "medium" proteins is another example of degree, and precision. In agreement with reviewer 2 I feel a concrete example of a 150 residue protein would do wonders for

credibility. The cited example of FimC shows only the very sensitive H α/β C α/β (CO)NHN experiment. The new Figure S20 purportedly showing the water suppression with a single transient also shows me that the 22kDa protein example is either (a) totally unfolded or (b) only the mobile regions are detected; the total Hn dispersion appears to be < 0.8ppm. I should also add that S20 (A) appears to *have* convolution applied at the water signal because the noise level at the center frequency *is* suppressed. The Figure does indeed show tolerably small signal for the residual water with one transient but it also shows me other details!

The response "we do not see that 4C,4E,4F,5A,5B and 5C are (within the signal-to-noise) out of phase" was subject to a blind test in my own laboratory and two researchers identified phase problems *and* the artefact at 73ppm in 4F. While agreeing on the problem of phase in 4B no explanation is offered, though there seems to have been some misunderstanding since the second part of the question "particularly why are HACA affected differently from HBCB" referred to the same experiment rather than a different pulse scheme as proposed by the authors.

The S/N in Figures S14 to S19 are integers - which is highly unlikely.

Still missing from the paper is the inclusion of time to process data-sets, and to reliably peak-pick them for AutoAssign. Certainly sorting of clean lists takes only a few CPU seconds but the expert intervention in peak-picking the input is rate limiting. The phase errors alluded to above, coupled with truncation effects and overlap evident in Fig 7, 8 and 9 make this a non-trivial step. There also needs to be a description of how AutoAssign extracts the individual shifts rather than referring to a Perl script.

There are many other details that need addressing but since this will be the subject of another reviewer, and after revision for another journal it does not seem useful to pursue further.

EXHIBIT 14

Dr. T. Szyperski
Associate Professor of Chemistry and Biochemistry
Department of Chemistry, 816 Natural Sciences Complex
University at Buffalo, The State University of New York, Buffalo, NY, 14260
Phone: 716-645-6800 ext2245, Fax: 716-645-7338
E-mail: szypersk@acsu.buffalo.edu

Dr. F. Ann Walker
Associate Editor, Journal of the American Chemical Society
The University of Arizona
Department of Chemistry
1306 E. University Blvd
Tucson,
Arizona 85721-0041

ja016178i: "Reduced-dimensionality NMR spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis", by T. Szyperski, D. Yeh, D. Sukumaran, H. Moseley and G. Montelione.

Dear Prof. Walker,

thank you very much for your fax of August 14, 2001. I am amazed to read the comments of reviewer 1 with respect to our revisions, and I don't know what has lead his response here.

I have never in my career replied to a final decision of an editor, but I think that this case deserves further attention. I would be grateful if you could look into this response of the reviewer in more detail in order to judge on the appropriateness.

(A) 1st para: The stick diagram.

To refer to the missing second b-proton in the stick figure for the HabCabCONHN experiment is unfair: for the b-protons only a single peak pair is shown throughout (evidently, for simplicity). Why does that become a point here? How to present these details is, after all, a matter of taste, and not something that should decide on the publication of a paper in JACS.

(B) 2nd para: size limitations.

The key point is that 3D RD NMR is as sensitive as 4D NMR, and the suitability of 4D NMR is amply demonstrated in literature. The HabCab(CO)NHN experiment is not the most sensitive.

The reviewer is simply wrong here.

(C) Quotation of the 21 kDa protein FimC.

The quotation of 21 kDa FimC with respect to the HabCab(CO)NHN experiment also shows the applicability of the entire suite of experiments. The reviewer states that this shows only that it works for the very sensitive HabCab(CO)NHN experiment. In fact, Figure 3 shows that the sensitivity of HabCab(CO)NHN experiment is in the "middle range". The reviewer is simply wrong here.

(D) Water suppression.

The FID of Fig S20 shows the signal of a nicely folded, purely alpha-helical protein with very low 1HN dispersion (I am ready to send a 2D [15N,1H]-HSQC). But the issue here is not the protein itself, but the water suppression with a single scan, and here the reviewer agrees that the FID for a single scan shows a tolerably small signal! - that is all what this figure in the supplementary material serves for (a marginal issue for the paper). We show what needs to be shown in this 20th figure of the supplementary material, and this reviewer is criticising us without need.

(E) Phase problems.

Please judge yourself with respect to the criticism of phase errors and artefacts (Figure 4). The attitude of the reviewer is non-constructive here, since all NMR spectra show non-ideal features. Neither automated nor manual analysis was impeded in any way.

(F) S/N distributions.

The integers of Figures S14-S19 are due to the fact, that all residues showing a signal-to-noise ratio of say, between 4.0 and 5.0, are plotted at 4.5. The number of residues is necessarily an integer. This is how many labs plot S/N distributions, and it is perfectly unclear to me how one can criticize this presentation.

(G) Extension of the paper for peak picking etc. We cannot at the same time shorten the Ms. and include new topics that are described in other papers (e.g. what we describe in a JBNMR paper, ref. 7c). **Please note that the reviewer has received Ref. 7c as a hard copy when reviewing this paper.** The points he criticized are amply described in this paper being submitted to a more specialized journal.

Overall, I consider the comments of this reviewer as unfair, and I would be more than grateful if you could consider a possibility to receive the opinion of the second reviewer on the revised Ms.

Please note, that the suite of RD experiment presented in this publication is received with great enthusiasm by the other groups of the NMR section of the Northeast Structural Genomics Consortium (Laboratories of M. Kennedy, C. Arrowsmith, G. Montelione). The RD NMR experiments are "used in the field" with success.....as described.

Please let me again emphasize that I have never in my career replied to a final decision of an editor, but I really think that the substance of what this reviewer is presenting here does not at all challenge the outstanding value of RD NMR for high-throughput efforts, or what we are pointing at in this paper.

Thank you for your support,
Yours sincerely,

Thomas Szyperski

EXHIBIT 15

Journal of the
American Chemical Society

PUBLISHED BY
THE AMERICAN CHEMICAL SOCIETY

Department of Chemistry
University of Arizona
1306 E. University Blvd.
Tucson, Arizona 85721-0041
Phone: (520) 626-9309
Fax: (520) 626-9300
Internet: jacs@u.arizona.edu

Dr. F. Ann Walker

September 6, 2001

By mail and fax 716 645 7338

[redacted]
Chemistry and Biochemistry
State University of New York at Buffalo
Department of Chemistry
816 Natural Sciences Complex
Buffalo, NY 14260

Ms No.: JA016178I Article
Title: "Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis"

Dear Dr. Szyperski:

Your manuscript was sent back to Reviewer 1, whose comments are again enclosed. Because of your desire for another review, I sought the advice of another reviewer, an independent expert in the field. Based upon this reviewer's evaluation of the manuscript and the previous reviews, I conclude that the manuscript is not appropriate for publication in the Journal. Therefore, I find that I must again unfortunately reject the paper and recommend that you heed the comments of all of the reviewers in preparing a revised manuscript for submission to another journal.

Thank you for submitting this manuscript to *J. Am. Chem. Soc.*

Sincerely yours,

F. Ann Walker
Associate Editor

Reviewer 2
Review for ja016178i (revised manuscript)
The revised manuscript of Szyperski et al. has significantly improved with respect to the original version by deleting much of the idealised projections to future achievements. This manuscript thus merits publication in a journal with a readership largely interested in the practical details of biomolecular NMR. The question remains whether it merits publication in the Journal of the American Chemical Society. Triple resonance experiments as well as reduced dimensionality spectroscopy are well known concepts widely used for the resonance assignment of proteins and nucleic acids. The detailed description of eleven RD triple-resonance experiments is certainly not of much interest to the broad readership of JACS. What would be of interest to many readers of JACS, especially molecular biologists interested in NMR as a tool to resolve structures, study molecular interfaces, etc., is the experimental proof that a certain set of NMR experiments combined with an automated assignment protocol will yield rapid assignment for a wide range of proteins. To achieve this goal, the proposed concept has to be applied to different (at least two or three) molecular systems in different molecular weight ranges and eventually using different isotope labels (e.g. partial deuteration). It is not sufficient to apply the method to a single small protein and add a reference to a couple of other proteins studied using the same strategy without providing any experimental evidence. By the way the spectrum of Fig. S20 also looks to me to correspond to a rather unstructured protein. Thus at the current state the proposed strategy is just one of many different possible strategies. I am actually using a different one based on another set of experiments, and neither me nor T. Szyperski and coworkers have proofed so far that their strategy is the more efficient one.
The scientific content does not gain much by putting the work in the context of structural genomics. A valuable strategy to speed up the assignment process will be of interest to any NMR spectroscopist working on proteins not only those involved in structural genomics. On the other side the work would certainly benefit from a more thorough analysis of other time limiting factors like data processing, peak picking, and automated assignment.
In agreement with Reviewer 1 I believe that the manuscript lacks sufficient experimental proof that the conclusions are valid for a large range of molecular systems, although some of the more specific points raised by Reviewer 1 can be argued to be rather subjecti ve and of minor importance.
In conclusion, I suggest that the major part of the manuscript containing the detailed analysis of the RD experiments and their application to the Z-domain should be published without significant changes in a more specialised NMR journal such as J. Biomol. NMR, and that the authors eventually resubmit at a later date a manuscript which discusses in more experimental detail their assignment strategy with respect to other methods for the example of representative protein systems without the need of a detailed description of NMR pulse sequences.

Reviewer
4

Journal of the American Chemical Society

PUBLISHED BY THE AMERICAN CHEMICAL SOCIETY

PLEASE RETURN: 1) by mail in the enclosed envelope, 2) by fax to (520) 626-9300, 3) e-mail your comments to jacs@u.arizona.edu. OR 4) submit your review on-line at http://pubs.acs.org/journals/jacsat/jacs_inst.html.acs.org/journals/jacsat/msreview.html NOTE: If you choose to fax this review or respond to the Associate Editor by email, fax, or online, you do NOT need to return the original review form.

ELECTRONICALLY SUBMITTED MANUSCRIPT **REVISED**

JA01 6178 I-25-a-117 Rec'd: MAY 10, 2001
Thomas Szyperski, Deok C. Yeh, Dinesh K. Sukumaran, Hunter Moseley, Gaetano T. Montelione

Reduced-dimensionality NMR Spectroscopy for High-Throughput Protein Resonance Assignment: Implementation and Automated Analysis

Recommendations:

- ☐ Publish without change
- ☐ Publish after minor revision
- ☐ Publish after major revision
- ☐ Do not publish
- ☒ Publish elsewhere-where? J. Magn. Reson. J. Biomol. NMR

Summary Rating:

Is the manuscript likely to be of interest to the broad readership?	☐ Yes	☒ No
Are the conclusions adequately supported by the data presented?	☐ Yes	☒ No
Are the literature references appropriate and correct?	☒ Yes	☐ No
Does the nomenclature used conform with accepted practice?	☒ Yes	☐ No
Are hazardous procedures clearly defined as such?	☒ Yes	☐ No

Comments:

Please be as specific as possible if revision by the authors is recommended. Indicate specifically whether descriptions of methods, tables of data, etc. should be reduced or eliminated with the understanding that they would be available to specialists as Supporting Information (SI) on the web.

The corresponding author of this manuscript disputes the Editor's decision rejecting the paper for publication in J.A.C.S. and recommending submission to a more specialized journal.

After reading the paper and associated correspondence, I agree with the Editor's decision. Both reviewers were initially of the opinion that the paper was not suitable for J.A.C.S., and the second review by Reviewer 1 reiterated this position. It seems to me that the points raised in the final correspondence from the author (disputing the Editor's decision) are not really the most important factors in the decision, but constitute more of a philosophical difference between the author and the reviewer. The main issue is that the subject of the paper is too specialized for the general readership of J.A.C.S. Publication in J.Biomol. NMR or even J.Magn.Reson. would be a much more appropriate option for this paper, since the target audience is actually a very small subset (those interested in rapid and automated NMR resonance assignments of proteins) of a small subset (those interested in NMR assignments of proteins) of a subset (those interested in biological NMR) of readers of J.A.C.S.

EXHIBIT 16

J|A|C|S
ARTICLES
Published on Web 08/18/2005

High-Resolution Iterative Frequency Identification for NMR as a General Strategy for Multidimensional Data Collection

Hamid R. Eghbalnia,*[,†,#,^] Arash Bahrami,[†,‡,§] Marco Tonelli,[†,‡,#] Klaas Hallenga,[†,‡,#] and John L. Markley[†,‡,§,#]

Contribution from the National Magnetic Resonance Facility at Madison, Center for Eukaryotic Structural Genomics, Graduate Program in Biophysics, Biochemistry Department, and Mathematics Department, University of Wisconsin—Madison, Madison, Wisconsin 53706

Received January 13, 2005; E-mail: eghbalni@nmrfam.wisc.edu

Abstract: We describe a novel approach to the rapid collection and processing of multidimensional NMR data: "high-resolution iterative frequency identification for NMR" (HIFI–NMR). As with other reduced dimensionality approaches, HIFI–NMR collects *n*-dimensional data as a set of two-dimensional (2D) planes. The HIFI–NMR algorithm incorporates several innovative features. (1) Following the initial collection of two orthogonal 2D planes, tilted planes are selected adaptively, one-by-one. (2) Spectral space is analyzed in a rigorous statistical manner. (3) An online algorithm maintains a model that provides a probabilistic representation of the three-dimensional (3D) peak positions, derives the optimal angle for the next plane to be collected, and stops data collection when the addition of another plane would not improve the data model. (4) A robust statistical algorithm extracts information from the plane projections and is used to drive data collection. (5) Peak lists with associated probabilities are generated directly, without total reconstruction of the 3D spectrum; these are ready for use in subsequent assignment or structure determination steps. As a proof of principle, we have tested the approach with 3D triple-resonance experiments of the kind used to assign protein backbone and side-chain resonances. Peaks extracted automatically by HIFI–NMR, for both small and larger proteins, included ~98% of real peaks obtained from control experiments in which data were collected by conventional 3D methods. HIFI–NMR required about one-tenth the time for data collection and avoided subsequent data processing and peak-picking. The approach can be implemented on commercial NMR spectrometers and is extensible to higher-dimensional NMR.

Introduction

The acquisition of multidimensional spectra normally is required for NMR investigations of biological macromolecules. To maintain a given level of digital resolution, the number of free induction decays (FIDs) that have to be recorded grows exponentially with the number of dimensions. In addition, relaxation during polarization-transfer steps leads to sensitivity losses as dimensions are added. As a result, data collection times for higher-dimensional spectra can be very long, and acquisitions frequently are limited by the stability of the biological sample. The usual practical compromise it to collect smaller data sets, which leads to lower resolution.[1]

Among the methods that have been proposed for fast data collection, one of the most successful is the reduced dimensionality (RD) approach.[2] Reduced dimensionality alleviates some of the difficulties of *n*-dimensional NMR by combining information from different evolution periods into a single dimension. One approach has been to create $^{15}N-^{13}C$ double- and zero-quantum coherence in a single evolution time.[3] Similarly, two-dimensional (2D) versions of HNCA and HNCO triple-resonance experiments, called MQ–HNCA and MQ–HNCO, have been developed.[4]

In more recent RD experiments, the dimensionality is reduced by monitoring the chemical shift evolution of two or more nuclei simultaneously as single-quantum coherences in a single indirect domain. For example, three-dimensional (3D) experiments can be recorded as 2D planes in which the two indirect chemical shifts are encoded in the second dimension. Although losses resulting from additional polarization transfers during the evolution periods cannot be avoided in RD experiments,[5] resolution and sensitivity gains can be realized.

A number of promising RD techniques have been described recently. These include the G-matrix Fourier transform (GFT)[6,7] and time-proportional phase incrementation (TPPI)[8] methods,

† National Magnetic Resonance Facility at Madison.
‡ Center for Eukaryotic Structural Genomics.
§ Graduate Program in Biophysics.
# Biochemistry Department.
^ Mathematics Department.

(1) Wider, G. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 193–275.
(2) Szyperski, T.; Yeh, D. C.; Sukumaran, D. K.; Moseley, H. N.; Montelione, G. T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8009–8014.
(3) Szyperski, T.; Wider, G.; Bushweller, J. H.; Wuthrich, K. *J. Biomol. NMR* **1993**, *3*, 127–132.
(4) Simorre, J. P.; Brutscher, B.; Caffrey, M. S.; Marion, D. *J. Biomol. NMR* **1994**, *4*, 325–333.
(5) Sattler, M.; Schleucher, J.; Griesinger, C. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93–158.
(6) Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385–1393.
(7) Natterer. F. *The Mathematics of Computerized Tomography*; Wiley: New York, 1986.
(8) Ding, K.; Gronenborn, A. M. *J. Magn. Reson.* **2002**, *156*, 262–268.

10.1021/ja052120i CCC: $30.25

which differ in the strategy used for extracting frequencies that evolve simultaneously. In GFT, quadrature detection of all simultaneously evolving signals is carried out. For each increment, multiple FIDs are recorded that encode all the possible combinations of sine and cosine terms for the evolving frequencies. A matrix, called the G-matrix, is then applied to linearly combine the FIDs and extract all the various frequency terms. Traditional Fourier transformation then provides a 2D spectrum for each combination of the evolving chemical shifts. In the TPPI approach to RD, quadrature detection is performed sequentially on the signals of each of the simultaneously evolving types of nuclei. After the two data sets are Fourier transformed, signals appear with positive and negative frequency offsets. These two signals correspond to the two evolving chemical shifts. In TPPI, an artificially large resonance offset for the carrier frequency is introduced to avoid overlap. As a result, the peaks representing different frequency combinations are located in distinct regions of the single 2D spectrum. In the case of a 3D spectrum, GFT provides a series of 2D planes each carrying a different combination of the simultaneously evolving frequencies, whereas the TPPI method locates peaks for different frequency combinations in distinct regions of a single 2D spectrum. The larger number of FIDs needed in GFT to achieve quadrature detection of all the convolved frequencies is offset in the TPPI approach by the larger number of increments collected to achieve the same digital resolution for the wider spectral domain that needs to be covered.[9]

The strength of RD methods lies in the possibility of both reducing the collection time of high-dimensional spectra and increasing their digital resolution. However, at the digital resolution achieved, some 3D peaks may still be overlapped. In this case, additional information would be needed to make assignments. Automated peak assignments require adequate definition of peak positions.[10] Thus, it is important for RD methods to provide maximal resolution.

A different approach for fast multidimensional NMR data collection was recently presented by Freeman and Kupce.[11] In the specific case of 3D to 2D reduction, this method generalizes GFT from a fixed angle to several different angles. Peaks that overlap at a given angle frequently can be resolved at other angles (combinations of chemical shifts). Two-dimensional tilted planes are collected by simultaneously incrementing the evolution times for two indirect chemical shifts, with the projection angle being determined by the ratio of the two incrementation times. This approach has the potential of speeding up the data collection by recovering the peak positions in multidimensional spectra from a limited number of 2D tilted planes—potentially without sacrificing spectral resolution.

We present here a novel method for fast data collection and analysis. HIFI–NMR uses an iterative approach to recover the peaks in 3D spectra from two orthogonal planes plus a minimal number of 2D tilted planes collected one-by-one at optimal angles determined from analysis of the prior data collected. Data collection is terminated when the addition of data from another plane is predicted not to improve peak recovery. By collecting data at multiple angles and by focusing on peak identification instead of complete 3D reconstruction, HIFI–NMR circumvents complications encountered with the processing of data from other reduced dimensionality approaches.[12] We demonstrate here the application of the approach to NMR spectroscopy of proteins with the goal of recovering an optimal number of peaks in a minimal amount of spectrometer time from HIFI–NMR versions of 3D triple-resonance NMR experiments.

The underlying principle behind HIFI–NMR is to combine the high digital resolution afforded by 2D spectra, the ability of tilted-plane data collection to separate overlapped peaks, and the flexibility afforded by real-time analysis of the emerging pattern of peaks from prior data collection so as to adaptively determine the next tilted plane to be collected or, alternatively, to terminate data collection because the model of extracted peaks cannot be improved by the addition of another plane. Peaks are identified from high digital resolution 2D spectra rather than from the inferior resolution of reconstituted 3D (or higher-dimensional) spectra. The collection of variable tilted-plane data allows one to avoid peak overlaps that occur as the result of the well-known correlation between ^{1}H and ^{13}C chemical shifts in $^{1}H-^{13}C$ moieties in proteins. This poses a potential problem with data collected at a 45° tilt plane (the default angle for the GFT–RD or TPPI–RD approaches) but can be overcome by collecting data at multiple angles (Figure 1).

Experimental Section

Rationale. The standard approach in NMR spectroscopy has long been the reconstruction of signals from a sampled time series by appealing to theories for perfect reconstruction. In HIFI–NMR, we have chosen to recover the chemical shift frequencies only (partial reconstruction), rather than to reconstruct the spectra (perfect reconstruction). The reasons for this approach are both practical and theoretical. On the practical side, recovery of the frequencies of peaks with a degree of confidence is sufficient for the purpose of NMR peak assignments. On the theoretical side, perfect reconstruction from sampled data remains an active area of research into carefully devised and application-specific approaches. From our experience, in comparison with standard three-dimensional Fourier transform (3D FT) methods, to determine the frequencies of a comparable number of peaks HIFI–NMR data collection requires on the order of 5–15% of the sampled points. For perfect reconstruction from a comparable sparse set, one may consider using the methods of nonuniform sampling,[13] statistical reconstruction,[14] or tomographic reconstruction.[15] These approaches, in their respective standard settings, have to deal with challenges that arise from smearing, blurring, and false peaks. To our knowledge, no approach has been developed that is suitable for perfect reconstruction of NMR data from the category of sparsely sampled data obtained by tilted planes described here. We discuss further the theory of reconstruction and provide relevant references in the Supporting Information.

A more modest goal is to recover certain important features of the spectra in a way that is most consistent with the data collected. In multidimensional, multinuclear NMR experiments collected for the purposes of assignments, the key features to be recovered are the frequencies. By using an overcomplete dictionary of functions constructed from shifted and scaled copies of the standard *sinc* function, we have devised a signal representation that is sparse in a sense that can be precisely defined. In addition, we have developed an efficient algorithm for the estimation of peak positions and for the refinement of the estimates for this representation through stepwise data collection.

(9) Kozminski, W.; Zhukov, I. *J. Biomol. NMR* **2003**, *26*, 157–166.
(10) Olson, J. B., Jr.; Markley, J. L. *J. Biomol. NMR* **1994**, *4*, 385–410.
(11) Freeman, R.; Kupce, E. *J. Biomol. NMR* **2003**, *27*, 101–113.
(12) Tonelli, M.; Metha, D. P.; Eghbalnia, H.; Westler, W. M.; Markley, J. L. Presented at the 45th Experimental NMR Conference, Asilomar, Pacific Grove, CA, April 18–23, 2004, Abstract 347.
(13) Feichtinger, H. G.; Pandey, S. S. *J. Math. Anal. Appl.* **2003**, *279*, 380–397.
(14) Tenorio, L. *Siam Rev.* **2001**, *43*, 347–366.
(15) Faridani, A.; Ritman, E. L. *Inverse Problems* **2000**, *16*, 635–650.

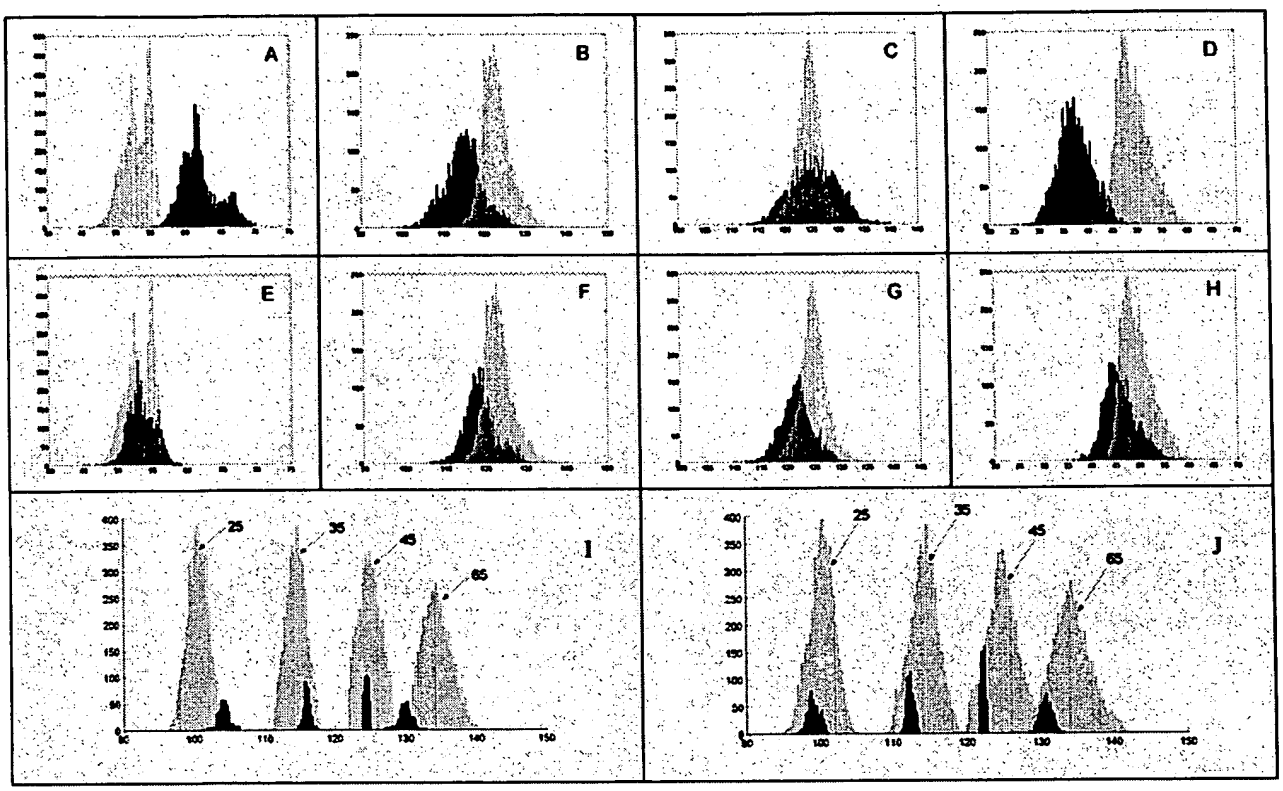


Figure 1. Illustration of the problem of *statistical indistinguishability* and its solution through data collection at multiple tilt angles. Shown are chemical shift distributions for Ala (yellow), Thr (red), and Asn (blue) in proteins (from BMRB, www.bmrb.wisc.edu on 03/10/2004). The first and second rows illustrate the statistics of combined $^{13}C^{\alpha}$ and ^{15}N chemical shift co-evolution corresponding to a plane at 45°. (Row 1) Comparison of the chemical shifts of Ala and Thr (a "best case") shows that they are partially distinguishable. (Row 2) Comparison of the chemical shifts of Ala and Asn (a more "typical" case) shows that they are statistically indistinguishable: (A, E) $^{13}C^{\alpha}$ chemical shifts; (B, F) ^{15}N chemical shifts; (C, G) sums of $^{13}C^{\alpha}$ and ^{15}N chemical shifts; (D, H) differences of $^{13}C^{\alpha}$ and ^{15}N chemical shifts. (Row 3) Effect of tilted-plane data collection (angle shown in the figure) on peak distinguishability for sums of $^{13}C^{\alpha}$ and ^{15}N chemical shifts: (I) full population of Ala chemical shifts compared with a subpopulation of Thr chemical shifts; (J) full population of Ala chemical shifts compared with a subpopulation of Asn chemical shifts. In each case (Thr and Asn), the subpopulation was selected to represent the 20% of chemical shifts closest to the mean. The relative movement of each subpopulation with respect to the Ala distribution illustrates the idea that distinguishability can always be achieved when multiple angles are considered.

We present the details of these issues, including the measure of optimality given the sparsity of coefficients in the representation, in a more mathematical setting in the Supporting Information. The basic idea is to iteratively refine the positions of peaks by optimally selecting planes that offer the "best evidence" for refining the positions of peaks that best explain the observed data. Our approach employs Bayesian methods to refine the peak positions after the collection and peak analysis of each new tilted plane.

For optimal performance, the HIFI−NMR strategy requires a number of components: pulse sequences adapted for tilted-plane data collection, a method for statistical analysis of spectral space, a robust automated peak-picking algorithm, an algorithm for analyzing prior data to choose the next-best plane, an algorithm for deciding when to cease data collection, and an algorithm for statistical, probabilistic analysis of peak positions (frequencies in each dimension) and their validity.

Data Collection. The 2D spectra required for extracting 3D resonances by HIFI−NMR are the two orthogonal planes (F_1−F_3, F_2−F_3), as well as a small number of carefully selected tilted planes (aF_1bF_2−F_3). The ratio between the dwell times of the two simultaneously evolving dimensions determines the angle of the resulting tilted plane. We set the spectral windows for F_1 and F_2 to certain values and multiply the corresponding dwell times for F_1 and F_2, respectively, by the sine and the cosine of the chosen tilt angle. (This is equivalent to dividing the spectral windows by the sine and cosine of the angle.) Whereas this approach has the advantage of presenting only a single recognized parameter for the user to set, it has the consequence of fixing the spectral window of each tilted plane.

To avoid aliased peaks in tilted planes, we note that the size x of a spectral window is given by $x = \Delta t_N/\Delta t_C = sw_C/sw_N$, and that the simultaneous evolution can be written as $\Sigma_i \cos\{(\omega_C \pm x\omega_N)\Delta t_C i\}$, where Δt_N and Δt_C are the sampling or dwell times, sw_C and sw_N are spectral windows, and the subscripts indicate the nitrogen and carbon frequency domains, respectively. The frequency $\omega_C \pm x\omega_N$ is thus sampled at a rate sw_C. For standard values sw_C^{MIN} and sw_N^{MIN}, respectively, $x = sw_C^{MIN}/sw_N^{MIN}$, and the maximum combined frequency can be no larger than twice the highest carbon frequency. In our experiments, the sampling/dwell times are multiplied by cos α and sin α; therefore, it is sufficient to make both spectral windows larger by a factor $2\cos(\pi/4)$. However, by taking advantage of the data in the orthogonal planes, we can find a more efficient spectral window by filtering out outlier peaks, which by definition are uniquely identifiable. As a result, in practice, a factor of 1.2 was found to be sufficient in nearly all cases. An alternative strategy of adjusting spectral windows on the basis of observed data and expected peak positions can be devised in order to further optimize data collection.

To achieve quadrature detection, we use the States method[16] independently for both F_1 and F_2, with the result that four FIDs are recorded for each increment in the spectrum. These four FIDs, which are collected interleaved for each increment, represent all possible combinations of real and imaginary points for the indirect dimensions.

(16) States, D. J.; Haberkorn, R. A.; Ruben, D. J. *J. Magn. Reson.* **1982**, *48*, 286−292.

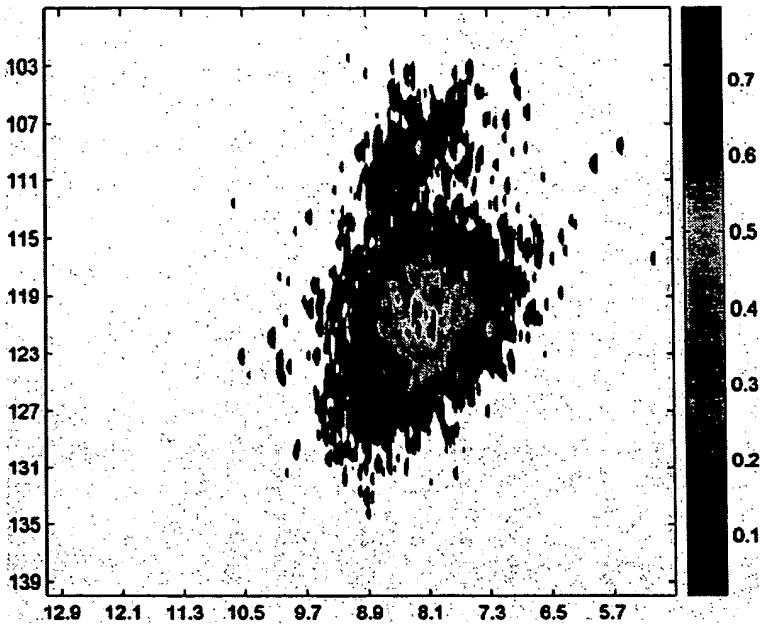


Figure 2. $^{1}H-^{15}N$ probability density plot (chemical shift priors) for the mouse protein Mm202773 generated from the sequence of the protein. The colors correspond to the probability scale on the right. The approach used in constructing chemical shift priors from data in BMRB will be described separately (A. Bahrami et al., to be published).

The States–TPPI modification for shifting axial peaks to the edge of spectra[17] is performed on only one of the simultaneously evolving dimensions.

Data are processed in a way similar to that described by Kozminsky and Zhukov.[9] The four FIDs are separated into two spectra, each containing the real and imaginary points for one of the simultaneously evolving dimensions (e.g. F_1), but only the real or the imaginary points of the second one (F_2). Fourier transformation of these two data sets results in two spectra, 90° out of phase, with duplicated peaks arising from the indistinguishable "+" and "−" F_2 frequencies.

The sum of these two spectra (after the phase of one of them is corrected by 90°) results in a spectrum that contains a single set of peaks with the frequencies of the two simultaneously evolving nuclei added together ("+ tilt" spectrum). The difference between the two spectra yields a spectrum containing the other set of peaks at the position of the subtracted frequencies ("− tilt" spectrum). The signal-to-noise ratio in either case would be equivalent to that of a single spectrum acquired over the sum of the data acquisition times of the two spectra.

In adapting 3D pulse sequences for HIFI–NMR, an important modification is made: the constant-time ^{15}N evolution that characterizes many triple-resonance experiments is converted into a constant–semiconstant-time evolution. This modification enables us to collect any desired number of points in the indirect dimension. This is critical to take full advantage of the ability of HIFI–NMR to resolve peaks by collecting 2D planes with more data points than would be reasonably possible with a full 3D data set. The ^{15}N evolution period behaves like a constant-time evolution, provided that the number of increments required can be contained within the constant-time delay, and it switches into a semiconstant-time evolution when a higher number of increments is required for adequate signal-to-noise. The excess time that cannot be accommodated by the constant-time delay is distributed uniformly throughout the ^{15}N increments. In testing pulse sequences incorporating this modified ^{15}N evolution period on both small (5 kD) and larger proteins (up to 20 kD) studied, we observed no significant distortion of the ^{15}N line shape, even when large numbers of increments were used.

(17) Marion, D.; Ikura, M.; Tschudin, R.; Bax, A. *J. Magn. Reson.* **1989**, *85*, 393–399.

In choosing the spectral window in the indirect dimension of tilted planes, it is important to realize that the simultaneous incrementation of both evolution times connects/interconnects the otherwise independent chemical shift evolutions. These evolutions can now be written as a product with a single time variable, $\cos(\omega_A t)\cos(x\omega_B t)$, where ω_A and ω_B denote the chemical shift evolution for the two simultaneously evolving nuclei, and x is the ratio between their time increments. By using standard trigonometric identities, this product can be rewritten as the sum of two terms, the first containing a weighted sum of the chemical shift evolutions and the second a weighted difference. This situation is similar to the double- and zero-quantum coherences created in the original RD experiments described by Szyperski[3] and Simorre,[4] and explains our observation that peaks in the indirect dimension of tilted planes can be aliased even though the individual simultaneous frequencies are within the range of the spectral window chosen for each of them. Thus, to avoid aliasing, it is necessary to choose a sufficiently wide spectral window in the tilted planes, as discussed above.

Statistical Organization of Spectral Space. We divide 3D spectral space into discrete voxels on the basis of prior information: the number of expected peaks, the nuclei involved and their expected chemical shift ranges, and the data collection parameters. Next, we obtain an estimate for the prior probability distribution of peak locations in the multidimensional spectral space. For this step, we combine information about the amino acid sequence of the protein and empirical as well as derived analytical distributions that are based on deposited chemical shifts (BMRB, www.bmrb.wisc.edu) by using a Markov Chain Monte Carlo (MCMC) inference method. The resulting joint probability distribution estimated in this step signifies the probability of a peak's existence in a specific voxel of the spectrum. In most cases, this distribution gives "strong evidence" to a few regions of spectral space and a relatively flat probability distribution in other regions (Figure 2). In regions of low probability, peaks generally are dispersed and readily detected. However, in regions of high probability, overlaps are likely, and so the algorithm collects additional planes whenever the model suggests evidence for ambiguity. We denote the prior distribution as $P_{pr}(X)$, where the elements of the vector $X = (x_1,x_2,x_3)$ represent the voxels in *three*-dimensional space.

Overview of Signal Recovery. The initial step in the HIFI approach is to collect data for the two orthogonal planes (0° and 90° planes). For experiments involving ^{1}H, ^{15}N, and ^{13}C dimensions (for example, HNCO, HNCACB), the 0° (^{13}C−^{1}H) and the 90° (^{15}N−^{1}H) planes are collected and peak-picked. Although any automated peak-picking routine can be used to identify peaks in each 2D plane collected, we use a new algorithm (PEANUT) to enhance the peaks and remove noise so as to improve the quality of peak-picking (H. R. Eghbalnia et al., manuscript in preparation). As in other manual or automated approaches to NMR peak-picking, each plane threshold is selected slightly below the noise level so as to recover as many peaks as possible for the given plane. The final output of HIFI−NMR is an objective probabilistic analysis of all detected peaks and their rankings, which can be combined with similar HIFI−NMR analyses from other experiments in order to provide an effective and unbiased global list of ranked peaks.

Because the ^{1}H dimension is common to the two planes, the possible candidates for peaks in 3D space can be generated by considering all combinatorial possibilities for peaks in the two planes that have the same ^{1}H chemical shift within a given tolerance. This generates many more candidates than actual peaks but will include most, but not all, real peaks. These candidate peaks restrain the prior probability space generated in the earlier step. Voxels that have no positive probability in the 0° and 90° planes are assigned a lower probability. These probabilities may increase later if evidence for these voxels is discovered in subsequent 2D planes. As additional data are obtained, the principle of statistical invariance requires us to recompile and reevaluate the evidence in the sample space. Heuristically, we must base the current prior distribution on all current data, regardless of the order in which they were obtained. The correct application of this idea is critical to the success of our approach. At the end of this stage, we have gained new information about the probability space consisting of candidate peaks; we designate this probability as $P_0(x)$.

Statistical Evaluation of Signals. For the moment, we assume that data for a new tilted plane at the optimal angle have been obtained. The projection of the current set of candidate peaks in 3D space on the tilted plane at angle θ is the map D,

$$\mathrm{R}^3 \xrightarrow{D} \mathrm{R}^2: \quad D(x_1,y_2,z_3) = (x_1, y_2\cos(\theta) + z_3\sin(\theta)) \qquad (1)$$

in which $\mathbf{X} = (x_1,x_2,x_3)$ are the Cartesian coordinates of a candidate peak. We first discretize the coordinates and obtain the associated voxel label. We next examine each candidate peak in turn, in light of the data from the new tilted plane and prior data, and update its probability on the basis of Bayes' rule.

To do this, we compare a candidate peak x that arises from the projection $x_\theta = D(x)$ with the set of automatically picked peaks a_1, a_2, a_3, ..., a_k (treated as distributions) in the latest tilted plane θ by defining the statistical event,

$$x\_observed\_in\_\theta \equiv (\exists a_i \ni d(x_\theta,a_i) < \delta_i) \qquad (2)$$

in which d represents the distance based on the divergence measure and δ_i is the allowable statistical deviation of peak a_i. In intuitive language, measuring the event $x\_observed\_in\_\theta$ is equivalent to asking whether or not the peak x has been observed in the new tilted plane designated by the angle θ. The use of divergence instead of Euclidean metric is desirable, because ^{1}H and the mixture of ^{15}N and ^{13}C have different tolerance values and the tolerance values vary from plane to plane.

By invoking Bayes' rule, we find that the probability of x being a "real peak", $P_0(x)$, is updated after observation of peaks in tilted plane θ by

$$P_1(x) \equiv P(x|x\_observed\_in\_\theta) = \frac{P(x\_observed\_in\_\theta|x)\,P_0(x)}{P(x\_observed\_in\_\theta)} \qquad (3)$$

or

$$P_1(x) \equiv P(x|x\_not\_observed\_in\_\theta) = \frac{P(x\_not\_observed\_in\_\theta|x)\,P_0(x)}{P(x\_not\_observed\_in\_\theta)} \qquad (4)$$

in which $x\_not\_observed\_in\_\theta$ is the complement event for $x\_observed\_in\_\theta$. By "event X" we mean that the candidate peak x is statistically distinguishable from noise and therefore is a real peak. Note that $P_0(x)$ is the prior probability that we have calculated in the previous stage. The conditional likelihood $P(x\_observed\_in\_\theta|x)$ represents the probability that the candidate peak x would be observed in tilted plane θ, given that it is a "real peak". The complement of this probability, $P(x\_not\_observed\_in\_\theta|x)$, is defined in the same way. The term in the denominator, $P(x\_observed\_in\_\theta)$ (or $P(x\_not\_observed\_in\_\theta)$), represents the probability of a voxel x in plane θ to be considered a peak (or not a peak), regardless of any other consideration. We consider this probability to be independent of the angle θ and obtainable only from our prior distribution. In cases where x is located in a crowded region, such that evidence for its existence may vary from plane to plane taken at different angles, we use empirical database information to determine an "independent estimate" by isolating the crowded region and by estimating the likelihood that our observation is false. The independent estimate enables us to approximate the probability $P(x\_observed\_in\_\theta)$.

The probabilities $P(x\_observed\_in\_\theta|x)$ and $P(x\_not\_observed\_in\_\theta|x)$ do not experience much angular variation and can be reasonably approximated by the empirical data.

Choice of the "Best Tilted Plane". The best tilted plane angle θ is selected according to the current probability distribution $P_0(x)$ and a mathematical method designed to obtain maximum information. We describe the method briefly here but point out that important technical steps (described in the Supporting Information) must be adhered to in order to obtain the optimal angle. The next-best tilted plane is the one that maximizes information about the positions of the peaks. To find this plane, we assume that the candidate peaks have been observed on a plane at angle θ and evaluate the information theoretic impact of this observation on the probability of the peaks. The measure of this influence is the divergence S_θ between the initial probability $P_0(x)$ (probability at the initial step, or step zero) and the predicted probability after the impact of the projection at angle θ has been taken into account, $Q_\theta(x)$:

$$S_\theta = -\sum Q_\theta \ln\frac{Q_\theta}{P_0} \qquad (5)$$

where the sum is taken over all voxels in the spectral space. The choice for our next plane optimizes this measure:

$$\theta_{\text{optimal}} = \arg\max_\theta (-S_\theta) \qquad (6)$$

Intuitively, by assuming a uniform prior over all plane selections, the maximum information method tends to choose as the next plane that with the highest dispersion of projected peaks. Two details (discussed in the Supporting Information) are worth noting here: (a) our above discussion of probabilities is in reference to the parameters of the model describing the spectra, and (b) a unique optimal angle may not exist.

After the first plane is chosen and the probabilities updated to $P_i(x)$ (probability at the ith step) by use of eqs 3 and 4, the next-best plane is again selected as described above. The probability updating rule after processing the ith plane θ_i is

$$P_i(x) \equiv P(x|x\_observed\_in\_\theta_i) = \frac{P(x\_observed\_in\_\theta_i|x)\,P_{i-1}(x)}{P(x\_observed\_in\_\theta_i)} \qquad (7)$$

or

$$P_i(x) \equiv P(x|x\_not\_observed\_in\_\theta_i) = \frac{P(x\_not\_observed\_in\_\theta_i|x)\, P_{i-1}(x)}{P(x\_not\_observed\_in\_\theta_i)} \quad (8)$$

This process is continued until no further information (or negligible information) can be obtained (i.e., the divergence is below a threshold). At this stage, n peaks with probabilities above a threshold will have been reported as the peak list. The value of n is estimated automatically from the number of amino acid residues in the protein, the type of experiment, and the probability distribution $P(x)$. The value of n can be adjusted in a subsequent postprocessing step to include additional expert information regarding the observed data.

Results and Discussion

Test of the HIFI–NMR Approach with 3D Spectra of Proteins. Prior to implementing the adaptive tilted-plane algorithm on an NMR spectrometer, we tested HIFI–NMR in a nonautomated environment. We collected orthogonal-plane and tilted-plane data (at either 5° or 10° intervals between 0° and 90°) according to the Freeman and Kupce RD method[11] for 3D triple-resonance experiments and then used the adaptive tilted-plane algorithm to select which of these in turn would be the next tilted plane for the peak-picking and analysis engine of HIFI–NMR. The proteins we used as test cases were ones that had been studied thoroughly by conventional methods. This enabled us to compare peaks "identified" by HIFI–NMR with those "identified" by manual peak-picking of conventional 3D data sets and to classify these peaks as "correct" (i.e., in agreement with all available data) or as "noise" (identified peaks that did not correspond to an assignment).

Data Sets Collected. To test the efficacy and accuracy of our approach, we colleted a total of eight data sets from five proteins: brazzein (54 residues), ubiquitin (76 residues), mouse protein Mm202773 (101 residues), *Anabaena variabilis* flavodoxin (179 residues), and Prp24_12 (166 residues). All proteins were labeled uniformly with carbon-13 and nitrogen-15.

We show results here for three experiments (HNCO, HNCACB, and CBCA(CO)NH) typically used in determining protein backbone assignments. As controls, we collected parallel data sets by the standard 3D data collection approaches, processed these by 3D FT, and handpicked the peaks for comparative analysis. For proteins other than Prp24_12, peak lists corresponding to solved NMR structures were used to validate peaks as "real"; in the case of Prp24_12, which had no solved structure, the only basis of comparison was handpicked 3D FT data.

Typically, for each 2D plane, four transients were accumulated for each FID (with the exception of the HNCACB planes of the mouse protein, in which eight scans were needed to observe all the intra- and interresidue cross-peaks). Given our modified ^{15}N evolution period (vide supra), 80 and 128 increments were collected in the indirect dimension of each plane (orthogonal or tilted) for HNCO and HNCACB experiments, respectively. However, in the CBCACONH pulse sequence, the constant-time evolution in the ^{13}C dimension limited the number of increments that could be recorded (for an 80 ppm spectral window) to 71. The orthogonal planes were recorded first, with two FIDs for each increment, to allow for quadrature detection in the indirect dimension, according to the States method. As described above, for each tilted plane, each FID was collected four times in order achieve quadrature detection for both simultaneously evolving nuclei. After processing, two planes were obtained for each chosen angle: those corresponding to the "+" and "−" tilt geometries.

In each experiment, we phased the initially collected plane and used the phasing parameters to phase all subsequent planes automatically. We found that specification of an approximate value for the noise level (as derived from the phasing step) was helpful in improving the efficiency of the plane selection algorithm but did not alter its predictions.

Comparison of Peaks Identified by HIFI–NMR to Those from the Controls. Ubiquitin proved not to be a challenging protein in that its peaks are nearly ideally distributed. For example, HIFI–NMR achieved the identification of all correct peaks in the 3D HNCO spectrum after the selection of a single extra plane beyond the two orthogonal planes. In fact, by experimentation we found that peak recovery was not particularly sensitive to the choice of angle for this third plane. For this reason, we do not discuss the ubiquitin data further here.

In the case of CBCA(CO)NH HIFI–NMR data collected for the protein brazzein (Table 1, panel A), we recovered the full 3D data with two orthogonal planes plus two tilted planes chosen sequentially by the HIFI–NMR engine. The first "best tilted plane" was predicted as 41°, and that actually used was the pre-collected plane at 40°. At that stage, the signal recognition module identified 103 peaks (including 82 of 87 correct peaks). The next-best plane was predicted as 34°, and that actually used was the pre-collected plane at 35°. Following analysis of these data, 116 peaks were recognized (including all 87 of the correct peaks).

The algorithm identified this as the stopping place, but as a test, a series of five additional best tilted planes were collected (at 50°, 20°, 25°, 60°, and 70°), and the results were analyzed sequentially (Table 1, panel A). The addition of these tilted planes yielded no more correct peaks and only served to increase the number of noise peaks. By comparison, the 3D FT control yielded 111 handpicked peaks (including all 87 correct). With a random selection of the pre-selected tilted planes (no next-best plane selection), four tilted planes were required on average to recover all 87 peaks (results not shown).

In the case of the 3D CBCA(CO)NH HIFI–NMR data collected for mouse protein Mm202773 (Table 1, panel B), the ideal first tilted plane was predicted as 51° (50° used), and three additional planes (55°, 40°, and 70°) were needed to reach the identified stopping point at which 241 peaks were detected (including 177 correct). The control 3D FT experiment identified 223 peaks (including 178 correct) (Figure 3). In this case, the addition of the next-best tilted plane beyond the predicted stopping point served to decrease the number of noise peaks from 64 to 45, but additional tilted planes increased the number of noise peaks without improving the number of correct peaks.

We collected two additional HIFI–NMR data sets (HNCACB, HNCO) for mouse protein Mm202773. The results (Table 1, panels C and D) clearly show that the number of tilted planes needed for optimal signal identification and their angles are dependent on the experiment type. Seven tilted planes were required to reach the stopping point in the HNCACB experiment, whereas only three were required in the HNCO experiment. For the HNCO experiment, the three sequentially chosen ideal next tilted planes were at 49°, 37°, and 9° (50°, 35°, 10°

Table 1. Comparison of the Number of Peaks Extracted Manually from Three Triple-Resonance Experiments (Used as Control) with Those Extracted Automatically by HIFI–NMR after the Collection of the Number of Tilted Planes Specified (1–7)[a]

Panel A CBCA(CO)NH – brazzein – 54 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		40°	35°	50°	20°	25°	60°	70°
Peaks identified	111	103	116	117	118	120	119	118
Correct peaks	87	82	87	87	87	87	87	87
Noise peaks	24	21	29	30	31	33	32	31
Panel B CBCA(CO)NH – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		50°	55°	40°	70°	30°	65°	20°
Peaks identified	223	215	201	225	241	222	232	241
Correct peaks	178	171	174	176	177	177	177	177
Noise peaks	45	44	27	49	64	45	55	64
Panel C HNCACB – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		20°	10°	30°	40°	50°	60°	70°
Peaks identified	387	425	371	403	427	439	419	472
Correct peaks	335	303	308	318	321	324	325	329
Noise peaks	52	122	63	85	106	115	94	143
Panel D HNCO – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		50°	35°	10°	70°	20°	25°	45°
Peaks identified	110	115	109	117	113	115	116	115
Correct peaks	91	90	91	92	91	91	91	91
Noise peaks	19	25	18	25	22	24	25	24
Panel E HNCO – Prp24_12 protein – 166 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5		
Tilted Plane		24°	38°	51°	71°	60°		
Peaks identified	221	182	180	166	175	185		
Correct peaks	136	117	122	133	134	135		
Noise peaks	85	65	58	33	41	50		

[a] Data were collected at 600 MHz on a Varian/Inova NMR spectrometer equipped with a cold probe. HIFI stops plane collection at columns marked with heavier lines and gray background.

used). By comparison, the first three optimal planes for HNCACB were located at 21°, 12°, and 32° (20°, 10°, 30° used). These results demonstrate that the optimal selection of planes plays an important role in the rapid identification of peaks.

The resolution of peaks in the optimal first plane for the HNCACB HIFI–NMR experiment (20°) was found to differ greatly from that of a plane acquired at 45°, the fixed tilt angle of the GFT–RD or TPPI–RD experiments (Figure 4). At 45°, many peaks canceled out, owing to overlaps of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals, whereas the plane at 20° showed much clearer peak separation.

Conclusions

Our results show that the HIFI–NMR approach is capable of automatically identifying 98–101% of the peaks shown to be correct in handpicked data from conventional 3D FT spectra (Table 1). The positions of the peaks picked automatically by HIFI–NMR were statistically indistinguishable from those determined manually (<0.04 ppm in ^{13}C and ^{15}N; <0.003 ppm in ^{1}H). However, as a percentage of total peaks, HIFI–NMR returned 4–14% more noise peaks than were handpicked from 3D FT data sets. The required time for HIFI–NMR data collection was typically on the order of one-tenth (never less than one-fourth) that for conventional 3D FT (Table 2). By providing probabilistic peak lists as output (peaks with corresponding signal likelihood and uncertainties in frequencies), HIFI–NMR obviates the need for lengthy postprocessing and peak-picking as needed for 3D FT or other RD approaches.

The mathematical technique and the algorithm introduced in HIFI–NMR represent important steps toward achieving more complete automation of biomolecular NMR spectroscopy. The nonautomated results shown here involved the processing of pre-collected tilted-plane data, which was made available to the algorithm one plane at a time according to the angle selected. Our results (Table 1) show that the angle of the ideal first plane depends on both the experiment and the protein and is rarely the same from one case to the next. Thus, adaptive sampling provides an efficient approach to data collection.

We recently have succeeded in integrating these tools with a commercial Varian NMR spectrometer so as to enable real-time collection by the HIFI–NMR approach. This has enabled us to achieve the predicted gains (Table 2) afforded by real-time adaptive tilted plane selection and to collect data at the actual tilt angle selected. Upon repeating HIFI–NMR data collection of the HNCO data for protein Mm202773, the tilted planes selected and used were 54°, 35°, and 71°. When requested to select an additional plane beyond the determined stopping point, a fourth plane was added at 8°. Interestingly, these angles are similar to those selected (49°, 37°, 9°, 70°) and used (50°, 35°, 10°, 70°) in the nonautomated trial, but they were chosen in a different order. In both cases, the algorithm called for the same number of tilted planes, and the numbers of correct peaks and noise peaks returned were identical.

With this integrated tool, we have collected data sets and extracted 3D peaks from two larger proteins (19–20 kDa). For

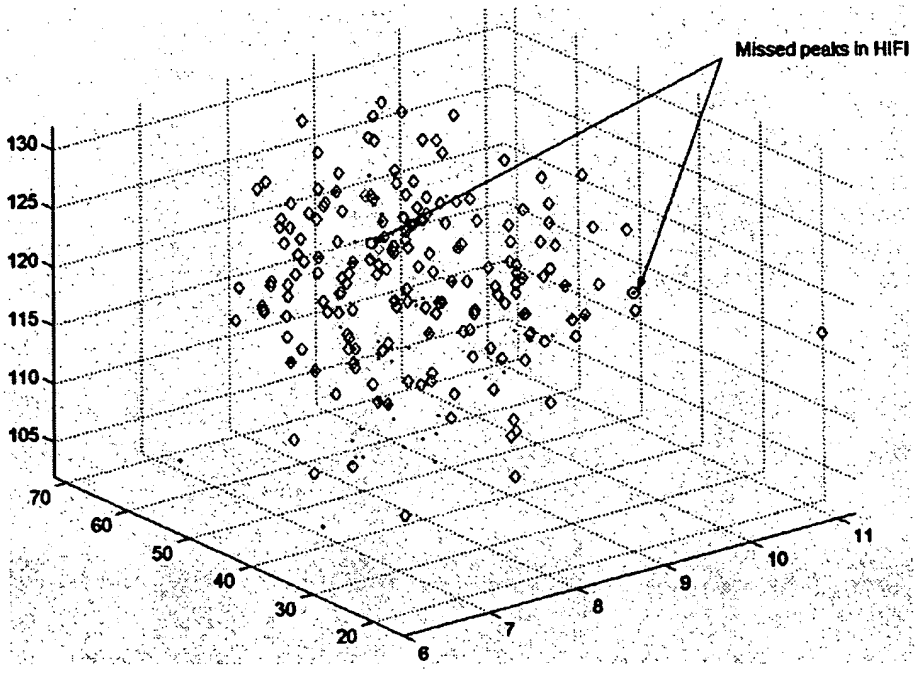


Figure 3. Schematic results from a HIFI–NMR version of the CBCA(CO)NH experiment. Displayed here are results generated from the two orthogonal planes plus three adaptively selected planes. The protein sample was mouse protein Mm202773. Red diamond symbols correspond to peaks identified by both HIFI–NMR and manual analysis of conventional 3D that correspond to real signals; green diamond symbols correspond to peaks identified by both HIFI–NMR and manual 3D that correspond to noise; blue dots correspond to peaks identified by manual 3D analysis but not by HIFI–NMR that correspond to real signals.

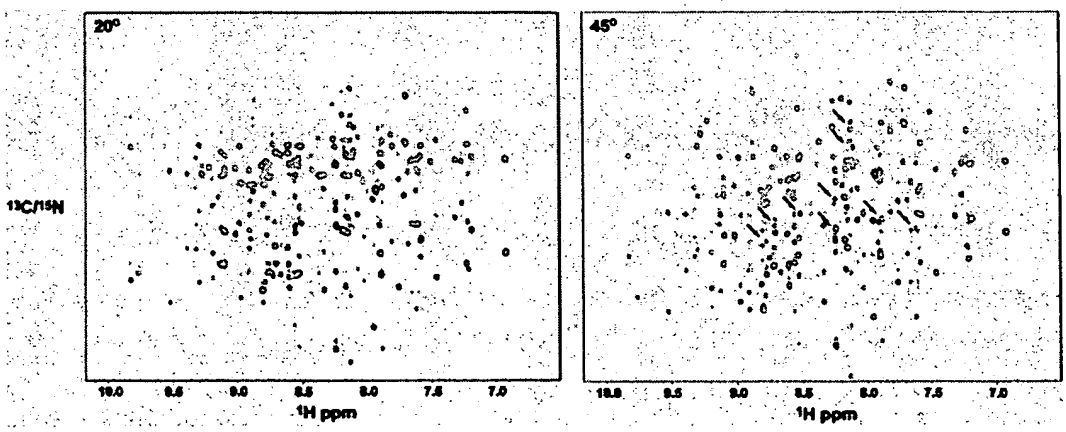


Figure 4. Two 2D tilted planes from the HNCACB HIFI–NMR spectrum of the mouse protein collected as described in the text. The 2D spectrum for the optimal first plane predicted by our algorithm at 20° is compared with the 2D spectrum for 45°. The arrows indicate the positions of signals that have been canceled because of spectral overlaps.

example, HNCO data from a perdeuterated sample of Prp24_12 (166 residues, 19.2 kDa) were collected and automatically processed from five planes in approximately 6.2 h to generate a 3D peak list with 135 entries. By comparison, 3D FT data collection for the same sample required 22 h (plus time for manual peak-picking) and resulted in 136 peaks, including all 135 identified by HIFI–NMR (Table 1, panel E). For flavodoxin (179 residues, 20 kDa), only four planes were needed to recover the peaks from HNCO or HNCOCA experiments. These results demonstrate that the current version of HIFI–NMR, when used for 3D NMR experiments of the type used in protein backbone assignments, can automatically recover more than 98% of the peaks found by conventional 3D FT in a fraction of the time.

The mathematical and computational tools proposed are sufficiently general to allow the combination of information from various sources into a single model. This will make it possible to improve the overall process through the integration of data

Table 2. Comparison of the Performance of HIFI–NMR with Conventional (Manual) Methods for Data Collection, Analysis, and Peak-Picking of Data from [U-^{13}C,U-^{15}N]-Mm2022773 (101-Residue Mouse Protein) Collected at 600 MHz on a Varian/Inova NMR Spectrometer Equipped with a Cold Probe

Experiment	CBCA(CO)NH		HNCO	
Data collected	HIFI	3D FT	HIFI	3D FT
Number of identified peaks:	215	223	117	110
Number of correct peaks	177	178	92	91
Total time required for data collection.	2 h[a]	22 h	1 h[a]	12 h

[a] Time for HIFI includes automatic generation of the peak list; this process is performed separately in 3D FT.

from multiple experiments. For example, by combining data from 3D experiments with common data planes (for example, CBCA(CO)NH, HNCACB, and HNCO), it should be possible to improve the discrimination between signal and noise. Most importantly, the direct output from HIFI–NMR is a peak list that can be used as input to automated assignment software packages.

The mathematical methods and associated tools developed here are part of a broader effort to achieve highly efficient and streamlined approaches for NMR structure determination and validation. A subset of these automation tools is available for public use at bija.nmrfam.wisc.edu.

Acknowledgment. This research was supported by Biomedical Research Technology Program, National Center for Research Resources, through NIH grant P41 RR02301, which supports the National Magnetic Resonance Facility at Madison. A.B. received partial support from the National Institute of General Medical Science's Protein Structure Initiative through NIH grant 1 P50 GM64598, which supports the Center for Eukaryotic Structural Genomics. During part of this work, H.E. was supported as a postdoctoral trainee by the National Library of Medicine under grant 5T15LM005359. We thank Eldon L. Ulrich and William M. Westler for advice and encouragement. This work made extensive use of the BioMagResBank and the Protein Data Bank.

Supporting Information Available: Mathematical details for selecting the next tilted plane and overcomplete representation, and an example illustrating the use of distributions on the parameter space for predictive density estimation of a normal model and how the parameters for intermediate densities need an enlarged space. This material is available free of charge via the Internet at http://pubs.acs.org.

JA052120I

EXHIBIT 17

Journal of Biomolecular NMR, **22**: 299–300, 2002.
KLUWER/ESCOM

Letter to the Editor: ^{1}H, ^{13}C, and ^{15}N resonance assignments and secondary structure of the PWI domain from SRm160 using Reduced Dimensionality NMR

Blair R. Szymczyna[a,c,*], Antonio Pineda-Lucena[a,c], Jeffrey L. Mills[b,c], Thomas Szyperski[b,c] & Cheryl H. Arrowsmith[a,c]

[a]*Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, 610 University Ave., Toronto, ON, Canada, M5G 2M9;* [b]*Department of Chemistry, University at Buffalo, The State University of New York, Buffalo, New York 14260, U.S.A.;* [c]*Northeast Structural Genomics Consortium*

Received 13 November 2001; Accepted 18 December 2001

Key words: pre-mRNA processing, PWI motif, reduced dimensionality NMR, resonance assignment

Biological context

SRm160 (the SR-related nuclear matrix associated protein of 160 kDa) belongs to a large group of pre-mRNA processing proteins which contain one or more domains rich in alternating serine/arginine residues (RS domains). SRm160 functions as a coactivator of both constitutive and exon-enhancer dependant splicing (Blencowe et al., 1998; Eldridge et al., 1999), and may play a role in the communication between splicing and 3′-end processing machinery (McCracken et al., 2002).

Here we report the nearly complete ^{1}H, ^{13}C and ^{15}N resonance assignments and secondary structure of a 12.5 kDa polypeptide containing the PWI motif of SRm160, of which the biological function is currently not known. The PWI motif, named after a highly conserved PWI tri-peptide at its N-terminal end, is highly conserved in homologs of SRm160 and other splicing or splicing-related proteins (Blencowe and Ouzounis, 1999). The assignments will serve as the basis for calculating the domain's three-dimensional solution structure, which will help elucidate its possible role in pre-mRNA processing. Assignment of the resonances was greatly facilitated by use of reduced dimensionality (RD) NMR methods.

Methods and experiments

A fragment of human SRm160 encoding amino acids 27-134, which includes the PWI motif, was cloned into the pET-15b expression vector (Novagen) and expressed in *E. coli* BL21-Gold (DE3) cells (Stratagene). ^{15}N, ^{13}C-labeled samples were synthesized in cells grown in standard M9 minimal media containing $^{15}NH_4Cl$ and ^{13}C-glucose and purified to homogeneity using Ni^{2+}-affinity chromatography. Samples were prepared in 25 mM phosphate buffer (pH = 7.0), 300 mM NaCl, 1 mM DTT, 10%D_2O/90%H_2O, to final protein concentrations between 1.3 and 1.6 mM.

NMR experiments were recorded at 25 °C on a Varian INOVA 600MHz spectrometer. Resolution of the spectra was increased using linear prediction in both the ^{15}N and ^{13}C dimensions. For data processing and analysis, the NMRPipe program package (Delaglio et al., 1995), and the SPSCAN (Glaser and Wüthrich) and XEASY (Bartels et al., 1995) programs were used.

Assignment of the protein backbone and side chain atoms was primarily accomplished using RD experiments with short acquisition times (Szyperski et al., 1998, submitted). 3D HNNCAHA [acquisition time: 7.5 h], HACA(CO)NHN [5.2 h], $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN [9.9 h], and conventional HNCACB [7.9 h] experiments were used to assign the backbone chemical shifts. Aliphatic side chain

*To whom correspondence should be addressed. E-mail: bszymcz@uhnres.utoronto.ca

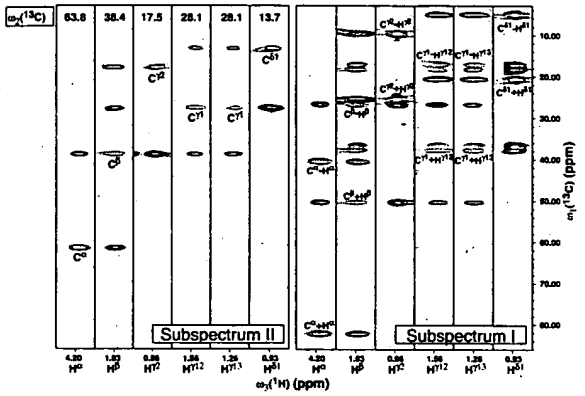


Figure 1. Contour plot of $[\omega_1(^{13}C), \omega_3(^1H)]$-strips taken from the two subspectra of the 3D HCCH-COSY experiment. Strips were taken at the sidechain ^{13}C chemical shifts of Ile 116 in $\omega_2(^{13}C)$ (indicated at top), and are centered about the corresponding ^1H chemical shifts in $\omega_3(^1H)$ (indicated at the bottom). Subspectrum II contains the central peaks, while Subspectrum I contains the RD NMR peak pairs, which are centered about the peaks in Subspectrum II. The observed correlations, indicated by solid lines, enabled identification of the spin system. Chemical shifts are relative to 2,2-dimethyl-2-silapentane-5-sulfonate.

chemical shifts were assigned predominantly using the 3D HCCH-COSY [8.9 h], and aided partially by heteronuclear resolved NOESY. Aromatic H^δ protons were linked to aliphatic H^β/C^β resonances using 2D $H^\beta C^\beta(C^\gamma C^\delta)HD$ [6.0 h], and the other aromatic chemical shifts were then obtained from 2D ^1H-TOCSY-HCH-COSY [7.0 h]. The total acquisition time for the entire set of RD experiments used was 44.5 h.

RD NMR experiments facilitated resonance assignment by resolving chemical shift degeneracies and providing additional correlations not observed in conventional 3D NMR experiments. Figure 1 shows signals observed for Ile116 in the two subspectra obtained from 3D HCCH-COSY, which was recorded with acquisition of central peaks (Szyperski et al., 1998). The resonances in subspectrum II are derived from ^{13}C magnetization, and provide the same information as the conventional (H)CCH-COSY. Subspectrum I is derived from ^1H magnetization, and exhibits pairs of peaks (doublets) centered about each ^{13}C frequency for a given CH^n moiety. As these pairs encode the chemical shifts of the associated hydrogens, the two 3D subspectra contain the same information as 4D HCCH-COSY. Consequently, the doublets may resolve degenerate carbon chemical shifts. For example, the two doublets associated with $C^{\gamma 1}$ in subspectrum I indicate that the two methylene hydrogen chemical shifts are not degenerate. Moreover, as the PWI domain of SRm160 is predominately α-helical, C^α chemical shift degeneracy would have impeded sequential resonance assignment. Using RD NMR techniques, splitting of the C^α signals by H^α chemical shifts resolved most of these degeneracies.

Extent of assignments and data deposition

The assignments of the PWI motif are virtually complete. With the exception of residual histidine tag amino acids at the N-terminus, the backbone is completely assigned and every backbone amide resonance is accounted for in the $[^{15}N,^1H]$-HSQC. The assignment of backbone H^α and C^α resonances is complete, and 98% of all side chain resonances are assigned. Secondary structure determination based on C^α/C^β, H^N, and H^α chemical shifts reveals four helical regions which include amino acids 45-60, 65-74, 81-91, and 94-113. Chemical shifts of the PWI motif are deposited in the BioMagResBank (http://www.bmrb.wisc.edu) under accession number 5162.

Acknowledgements

This work was supported by grants from the Canadian Institutes of Health Research, the National Science Foundation (MCB 0075773 to T.S.), and the National Institutes of Health (P50 GM62413-01 to C. A. and T.S.). We thank Dr Benjamin Blencowe for providing the plasmid containing SRm160 and for helpful discussions, Ying Lu for help with cloning and Kyoko Yap for providing a program to calculate the secondary structure.

References

Bartels, C., Xia, T., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

Blencowe, B.J. and Ouzounis, C.A. (1999) *Trends Biochem. Sci.*, **24**, 179–180.

Blencowe, B.J., Issner, R., Nickerson, J.A. and Sharp, P.A. (1998) *Genes Dev.*, **12**, 996–1009.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

Eldridge, A.G., Li, Y., Sharp, P.A. and Blencowe, B.J. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 6125–6130.

Glaser R.W. and Wüthrich, K., http://gaudi.molebio.uni-jena.de/~rwg/spscan.

McCracken, S., Lambermon, M. and Blencowe, B.J. (2002) *Mol. Cell. Biol.*, **22**, 148–160.

Szyperski, T., Banecki, B., Braun, D. and Glaser, R.W. (1998) *J. Biomol. NMR*, **11**, 387–405.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Mosely, H.N.B. and Montelione, G.T. (2001), submitted.

EXHIBIT 18

Journal of Biomolecular NMR **27**: 285–286, 2003.

Letter to the Editor: Resonance assignments for the hypothetical protein yggU from *Escherichia coli*

James M. Aramini[a], Jeffrey L. Mills[b], Rong Xiao[a], Thomas B. Acton[a], Maggie J. Wu[a], Thomas Szyperski[b] & Gaetano T. Montelione[a,c,*]

[a]*Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, and Northeast Structural Genomics Consortium, U.S.A.;* [b]*Department of Chemistry and Structural Biology, The State University of New York, Buffalo, NY 14260, and Northeast Structural Genomics Consortium, U.S.A.;* [c]*Department of Biochemistry, Robert Wood Johnson Medical School, UMDNJ, Piscataway, NJ 08854, U.S.A.*

Received 21 April 2003; Accepted 13 May 2003

Key words: alpha-beta protein, AutoAssign, reduced-dimensionality triple resonance NMR, structural genomics

Biological context

In recent years, several structural genomics initiatives have been established world-wide aimed at rapidly elucidating protein structures of functional and biological interest, as well as providing a more comprehensive picture of protein conformational space. The Northeast Structural Genomics Consortium (NESG) is particularly focused on clusters of eukaryotic domain families from several model organisms, including humans, and homologous proteins from bacteria and archaea (www.nesg.org).

NESG target ER14 is a 100-residue hypothetical protein, yggU, from *Escherichia coli* [Swiss-Pro ID, P52060]. ER14 is a highly soluble, basic (pI = 9.1) protein of unknown function. This protein shows some sequence identity (~26%) to a second NESG target from *Methanobacterium autotrophicum*, TT135, whose recent solution structure revealed a novel protein fold (Pineda-Lucena et al., 2002). In this note we report the nearly complete backbone and side chain ^{1}H, ^{13}C, and ^{15}N resonance assignments of ER14, determined by a combination of conventional triple-resonance and novel reduced-dimensionality NMR techniques (Szyperski et al., 2002).

*To whom correspondence should be addressed. E-mail: guy@cabm.rutgers.edu

Methods and experiments

Uniformly ^{13}C,^{15}N-enriched ER14 was cloned, expressed and purified following standard protocols used in our consortium. Briefly, the full-length gene (*yggU*) from *E. coli* was cloned into a pET21d (Novagen) derivative, yielding the plasmid pER14-21. The resulting ER14 open reading frame contains eight non-native residues at the C-terminus (LEHHHHHH) of the protein. *E. coli* BL21 (DE3) pMGK cells were transformed with pER14-21, and cultured in MJ minimal medium (Jansson et al., 1996) containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose. The final yield of pure *U*-^{13}C,^{15}N ER14 (>97% homogeneity by SDS-PAGE; 12.5 kDa by MALDI-TOF mass spectrometry) was ~10 mg l^{-1}.

Samples of *U*-^{13}C,^{15}N ER14 for NMR spectroscopy were prepared at a concentration of 1.0 mM in 95% H_2O/5% D_2O solution containing 20 mM MES, 50 mM NaCl, 5 mM DTT at pH 6.5. All NMR data were collected at 20 °C on four-channel Varian INOVA 500, 600, and 750 MHz NMR spectrometers, processed with NMRPipe 2.1 (Delaglio et al., 1995), and analyzed using SPARKY 3.106 (Goddard and Kneller, Univ. Calif., San Francisco). Proton chemical shifts were referenced to DSS, while ^{13}C and ^{15}N chemical shifts were referenced indirectly using the gyromagnetic ratios of ^{13}C:^{1}H (0.251449530) and ^{15}N:^{1}H (0.101329118), respectively. Backbone (H^N, H^α, N, C′, C^α) and C^β resonance assignments were made with AutoAssign

1.9 (Zimmerman et al., 1997; Moseley et al., 2001), using peak lists from 2D ^{1}H-^{15}N HSQC and 3D HNCO, HN(CO)CACB, HNCACB, HN(CO)CA, HNCA, HA(CA)NH, and HA(CACO)NH spectra, along with glycine-specific spectral information derived from a 2D (H^N-N plane) Gly-phased HA(CACO)NH (Montelione et al., 1999). Backbone assignments were also determined by manual analysis of backbone reduced-dimensionality (RD) experiments, $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN, HACA(CO)NHN, HNNCAHA, $H^{\alpha/\beta}$ $C^{\alpha/\beta}$COHA, and HNN<CO,CA> (Szyperski et al., 2002). Side chain aliphatic assignments were determined using 3D (H)CC(CO)NH-TOCSY, H(CCCO)NH-TOCSY, HCCH-COSY, and RD HCCH-COSY experiments. Side chain aromatics were assigned using 2D HBCB(CGCD)HD and H-TOCSY-HCH-COSY RD experiments and 3D ^{13}C-edited NOESY spectra. Individual Asn/Gln side chain amide protons (H(*E*) and H(*Z*)) of were assigned on the basis of NOE intensity in both 3D ^{15}N- and aliphatic ^{13}C-edited NOESY spectra as described previously (Montelione et al., 1984).

Extent of assignments and data deposition

Using the assignment strategy outlined above and neglecting the C-terminal tag, we obtained nearly complete backbone (C′: 98/100; C^{α}: 99/100; H^N-N: 92/93; H^{α}: 107/108) and side chain (C^{β}: 91/92; C^{γ}: 84/89; C^{δ}: 52/59; C^{ϵ}: 14/19; H^{β}: 150/152; H^{γ}: 123/125; H^{δ}: 78/81; H^{ϵ}: 37/48; N^{ϵ}: 1/4) assignments for ^{1}H, ^{13}C, and ^{15}N signals. The assigned ^{1}H-^{15}N HSQC spectrum of ER14 and AutoAssign connectivity map are shown in Figures 1A and B, respectively. Our Assignment Validation Suite (AVS) software (H.N.B. Moseley and G.T. Montelione, in preparation) revealed three unusually upfield shifted ^{1}H resonances, H^N of A44 (5.62 ppm; Figure 1A), an H^{β} of D28 (1.21 ppm), and an H^{γ} of Q62 (1.06 ppm). The ^{1}H, ^{13}C and ^{15}N chemical shift data have been deposited in the BioMagResBank (accession number 5596). These chemical shift data corroborate our initial tertiary structure of ER14, which demonstrates that ER14 is an α/β protein, featuring a long helix and a number of short β-stretches in a ββββαββα topology. We are currently refining the solution structure of ER14 using residual dipolar coupling and stereospecific assignment strategies.

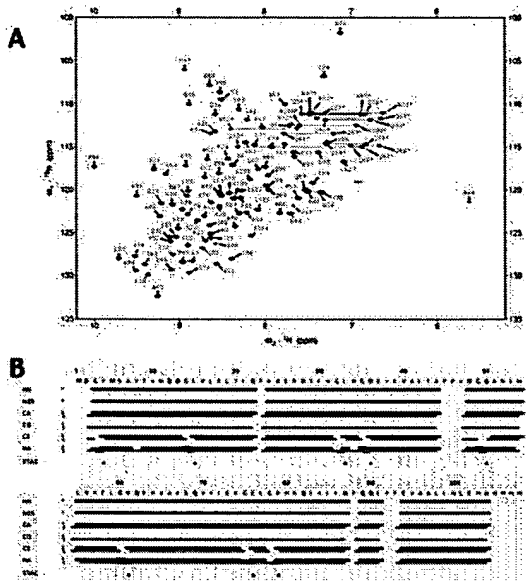


Figure 1. (A) Assigned ^{1}H-^{15}N HSQC spectrum of *E. coli* ER14, pH 6.5, 20 °C. Lines denote pairs of Asn/Gln side chain amide protons, each labeled with asterisks. (B) AutoAssign connectivity map for ER14. Intraresidue and sequential connectivities for the three-rung assignment strategy matching intraresidue and sequential H^{α}, C^{α}, and C^{β} resonances (Moseley et al., 2001) are shown as horizontal black and grey lines, respectively. Sequential spin-system typing assignment constraints (STACs), obtained from the Gly-phased HA(CACO)NH and used as overrides in AutoAssign, are represented by filled triangles.

Acknowledgement

This work was supported by a grant from the NIH Protein Structure Initiative (P50 GM62413) and the National Science Foundation (MCB 0075773) to T.S.

References

Delaglio, F. et al. (1995) *J. Biomol. NMR*, **6**, 277–293.
Jansson, M. et al. (1996) *J. Biomol. NMR*, **7**, 131–141.
Montelione, G.T. et al. (1984) *J. Am. Chem. Soc.*, **106**, 7946–7958.
Montelione, G.T. et al. (1999) In *Biological Magnetic Resonance*, Berliner, L.J. and Krishna, N.R. (Eds.), pp. 81–130.
Moseley, H.N.B. et al. (2001) *Meth. Enzymol.*, **339**, 91–108.
Pineda-Lucena, A. et al. (2002) *J. Biomol. NMR*, **22**, 291–294.
Szyperski, T. et al. (2002) *Proc. Natl. Acad. Sci USA*, **99**, 8009–8014.
Zimmerman, D.E. et al. (1997) *J. Mol. Biol.*, **269**, 592–610.

EXHIBIT 19

Journal of Biomolecular NMR **27**: 187–188, 2003.

Resonance assignments for the 21 kDa engineered fluorescein-binding lipocalin FluA

Gaohua Liu[a,*], Jeffrey L. Mills[a,*], Tracy A. Hess[a], Seho Kim[a,d], Jack J. Skalicky[a,e], Dinesh K. Sukumaran[a], Eriks Kupce[b], Arne Skerra[c] & Thomas Szyperski[a,**]

[a]*Department of Chemistry, University at Buffalo, The State University of New York, Buffalo, NY 14260, U.S.A.;* [b]*Varian NMR Systems, Oxfordshire OX8 1JN, U.K.;* [c]*Lehrstuhl für Biologische Chemie, Technische Universität München, 85350 Freising-Weihenstephan, Germany;* [d]*Present address: Department of Chemistry, Rutgers University, Piscataway, NJ 08854, U.S.A.;* [e]*Present address: National High Magnetic Field Laboratory, Tallahassee, FL 32310, U.S.A.*

Received 31 March 2003; Accepted 13 May 2003

Key words: anticalin, molecular recognition, protein design, reduced dimensionality NMR, resonance assignment

Biological context

The lipocalins form a large family of extracellular proteins which serve for transport and storage of secondary metabolites such as lipids, pheromones or prostaglandins (Flower, 1996). In spite of large diversity at the sequence level, lipocalins are structural homologues: a single eight-stranded antiparallel β-barrel with an attached α-helix forms the distinct 'lipocalin scaffold'. One end of the barrel is opened to the solvent and contains a ligand binding site. A set of four loops connecting consecutive strands confer specificity for ligand binding. Recently, the lipocalin scaffold was used to engineer proteins with tailored specificity for non-natural ligands. Such designed lipocalins can be considered as antibody mimics and were thus named 'anticalins' (for a review, see Skerra, 2000). Starting with the bilin-binding protein (BBP) from *P. brassicae*, the anticalin 'FluA' with binding affinity toward fluorescein was created using a combinatorial protein design approach. Compared to the amino acid sequence of BBP, FluA contains 20 point mutations and binds fluorescein with high specificity and affinity (Beste et al., 1999). As a step toward exploring the structural basis of molecular recognition by anticalins and lipocalins in general, we have overexpressed and purified several stable isotope labeled samples of FluA(R95K). Here we report the nearly complete ^{15}N, ^{1}H and ^{13}C resonance assignments.

Methods and experiments

FluA(R95K) protein containing a C-terminal *Strep*-tagII (Skerra and Schmidt, 2000) was overexpressed in *E. coli* KS474 using the plasmid pBBP21-FluA(R95K) essentially as described (Beste et al., 1999). Bacterial cell cultures were grown at 37 °C in (i) rich LB medium to express unlabeled FluA(R95K) (sample 1), in (ii) M9 minimal medium containing $^{13}C_6$-glucose and/or $^{15}NH_4Cl$ as sole carbon and nitrogen sources to produce either uniformly $^{13}C/^{15}N$-labeled (sample 2) or ^{15}N-labeled (sample 3) FluA(R95K), or in (iii) a M9 minimal medium containing $^{15}NH_4Cl$ in 30% H_2O/70% D_2O to synthesize ~50% deuterated, uniformly ^{15}N-labeled FluA(R95K) (sample 4). In addition, a fraction of sample 1 was lyophilized and dissolved in D_2O yielding sample 1b. Sample purity (>95%) and stable isotope labeling were verified by SDS-PAGE and MALDI-TOF mass spectrometry. All NMR samples were prepared at 0.7 mM protein concentration in 90% H_2O/10% D_2O (150 mM NaCl, 10 mM Na-PO_4, 0.2 mM EDTA, 50 mM benzamidine, pH = 6.4).

NMR spectra were recorded at 25 °C on Varian INOVA NMR spectrometers operating at ^{1}H resonance frequencies of 600, 750 or 900 MHz. Resonance assignments were obtained by combining (i) 2D [^{15}N,^{1}H]-TROSY (Pervushin et al.,

*Contributed equally to this work.
**To whom correspondence should be addressed. E-mail: szypersk@chem.buffalo.edu

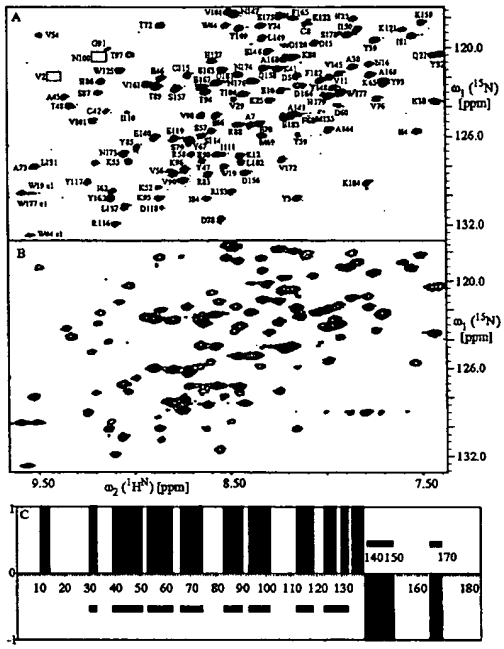


Figure 1. (A) 2D [^{15}N,^{1}H]-TROSY spectrum (Pervushin et al., 1997) recorded on a 900 MHz Varian INOVA spectrometer (25 °C) for uniformly ^{15}N-labeled and 50% deuterated FluA(R95K). The peaks are labeled using the one-letter code for amino acids. The peaks of Val 2 and Asn 100 are close to the noise level and their positions are indicated by boxes. (B) 2D [^{15}N,^{1}H]-HSQC spectrum recorded at 600 MHz with the same maximal evolution times as the spectrum shown in (A). (C) Chemical shift index (CSI) consensus plot (Wishart and Sykes, 1994) for identification of regular secondary structure elements. The eight β-strands forming the lipocalin β-barrel (black bars) as well as the α-helices (grey bars) are indicated.

1997; Figure 1) and a ^{15}N-resolved [^{1}H,^{1}H]-NOESY-[^{15}N,^{1}H]-TROSY acquired for sample 4 at 900 MHz, (ii) 2D [^{1}H,^{1}H]-TOCSY and NOESY acquired for both samples 1a and 1b at 900 MHz, (iii) 3D HNNCACB, CBCA(CO)NHN, HC(C)H COSY / TOCSY (Cavanagh et al., 1996) acquired for sample 2 at 600 or 750 MHz, (iv) reduced-dimensionality 3D H$^{\alpha/\beta}$C$^{\alpha/\beta}$(CO)NHN, 3D HNNCAHA, HACA(CO)NHN, 3D HCCH COSY, 2D HBCB(CGCD)HD, ^{1}H-TOCSY relayed HCH COSY (Szyperski et al., 1998; 2002) acquired for sample 2 at 600 and 750 MHz, and (v) 3D ^{13}C- and ^{15}N-resolved [^{1}H,^{1}H]-NOESY (Cavanagh et al., 1996) acquired, respectively, for samples 2 and 3 at 750 MHz.

Extent of assignments and data deposition

The combined use of double and triple resonance 3D spectra along with the 2D homonuclear spectra acquired at 900 MHz provided assignments (Figure 1) for 95% of the backbone and ^{13}C$^{\beta}$, and for 91% of the side chain chemical shifts of FluA(R95K). The measurement of ^{15}N spin relaxation parameters (Szyperski et al., 1993) revealed that FluA(R95K) reorients with a correlation time of ~10 ns at 25 °C (confirming that the protein is monomeric in solution). Thus, the use of TROSY (Pervushin et al., 1997) at 900 MHz dramatically improved the resolution of the 2D [^{15}N,^{1}H]-correlation map (Figures 1A,B), which facilitated the ^{15}N-^{1}H^N spin system identification. Based on the chemical shift data, the strands forming the lipocalin β-barrel can be readily identified (Figure 1C). FluA(R95K) possesses a rather large number of aromatic residues (5 Phe, 15 Tyr, 7 Trp, 7 His), and the resonance assignment of the aromatic rings greatly benefited from employment of the traditional homonuclear approach (Cavanagh et al., 1996) at 900 MHz. The ^{1}H, ^{13}C and ^{15}N chemical shift data have been deposited in the BioMagResBank database (accession number 5756).

Acknowledgements

This work was supported by the National Science Foundation (MCB 0075773 to T.S.) and the Fonds der Chemischen Industrie (grant to A.S.). A.S. thanks Josef Danzer for technical assistance in the production of isotope-labeled FluA(R95K).

References

Beste, G., Schmidt, F.S., Stibora, T. and Skerra, A. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 1898–1903.

Cavanagh, J., Fairbrother, W.J., Palmer III, A.G. and Skelton, N.J. (1996) *Protein NMR Spectroscopy*, Wiley, New York, NY.

Flower, D.R. (1996) *Biochem. J.*, **318**, 1–14.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 12366–12371.

Skerra, A. (2000) *Biochim. Biophys. Acta*, **1482**, 337–350.

Skerra, A. and Schmidt, T.G.M. (2000) *Meth. Enzymol.*, **326A**, 271–304.

Szyperski, T., Banecki, B., Braun, D. and Glaser, R.W. (1998) *J. Biomol. NMR*, **11**, 387–405.

Szyperski, T., Luginbühl, P., Otting, G., Güntert, P. and Wüthrich, K. (1993) *J. Biomol. NMR*, **3**, 151–164.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N.B. and Montelione, G.T. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 8009–8014.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

EXHIBIT 20

Journal of Biomolecular NMR 29: 549–550, 2004.

Letter to the Editor: Resonance assignments for the 18 kDa protein CC1736 from *Caulobacter crescentus*

Yang Shen[a], Hanudatta S. Atreya[a], Rong Xiao[b], Thomas B. Acton[b], Ritu Shastry[b], LiChung Ma[b], Gaetano T. Montelione[b,c] & Thomas Szyperski[a,*]

[a]*Department of Chemistry and Structural Biology, The State University of New York, Buffalo, NY 14260, U.S.A. and Northeast Structural Genomics Consortium;* [b]*Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, U.S.A. and Northeast Structural Genomics Consortium;* [c]*Department of Biochemistry, Robert Wood Johnson Medical School, UMDNJ, Piscataway, NJ 08854, U.S.A.*

Received 11 March 2004; Accepted 15 April 2004

Key words: alpha-beta protein, Northeast Structural Genomics Consortium, reduced-dimensionality NMR

Biological context

Structural genomics aims at making three-dimensional (3D) structural information readily available for each protein domain family in nature. In the Northeast Structural Genomics Consortium (NESGC; www.nesg.org), our efforts to solve three-dimensional protein structures focus on protein families encoded in eukaryotic genomes and their homologues in bacterial and archaeabacterial 'reagent' genomes (Wunderlich et al., 2004). Among these reagent organisms is *Caulobacter crescentus*, which is a proteobacterium known to differentiate and divide asymmetrically in each cell cycle (Laub et al., 2000). Owing to the small size of the genome containing only about 4000 genes, this organism has been selected as a model system to study cellular differentiation and cell cycle regulation (Laub et al., 2000). The recent sequencing of the genome of *C. crescentus* has fostered novel insights into its comparably complex cell cycle (Nierman et al., 2001). Here, we report nearly complete ^{1}H, ^{13}C and ^{15}N resonance assignments of a 148-residue protein of unknown function from *C. crescentus* that is encoded in gene CC1736 (NCBI ID: 155892; Swiss-Prot ID: Q9A7I7; NESGC target ID: CcR19; NESGC Rost-cluster ID: 17538).

Methods and experiments

Uniformly ^{13}C,^{15}N-enriched CC1736 was expressed and purified following standard protocols used in our structural genomics consortium. Briefly, the full-length CC1736 gene from *Caulobacter crescentus* was cloned into a pET21d (Novagen) derivative, yielding the plasmid pCcR19-21.1. The resulting open reading frame contains eight non-native residues at the C-terminus (LEHHHHHH) of the protein. *E. coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with pCcR19-21.1, and cultured in MJ9 minimal medium (Jansson et al., 1996) containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as the sole nitrogen and carbon sources. Initial growth was carried out at 37 °C until the optical density of the culture medium ($O.D_{600}$) reached 0.7 units. The incubation temperature was then decreased to 17 °C and protein expression was induced by the addition of IPTG (isopropyl-β-D-thiogalactopyranoside) at a final concentration of 1 mM. Following overnight incubation at 17 °C, cells were harvested by centrifugation and lysed by sonication. *U*-^{13}C, ^{15}N CC1736 was purified in a two step protocol consisting of Ni-NTA affinity column (Qiagen) and gel filtration column (HiLoad 26/60 Sephadex 75, Amersham Pharmacia Biotech) chromatography. The final yield of the protein (>97% homogeneity by SDS-PAGE; 17.7 kDa by MALDI-TOF mass spectrometry) was ~10 mg/L.

U-^{13}C,^{15}N CC1736 samples for NMR spectroscopy were prepared at a concentration of 1.1 mM in 95% H_2O/5% D_2O solution (20 mM MES, 100 mM NaCl, 5 mM $CaCl_2$, 0.02% NaN3, 10 mM DTT, pH 6.5) and placed in 5-mm Shigemi tubes. NMR spectra were collected at 25 °C on a Varian INOVA

*To whom correspondence should be addressed. E-mail: szypersk@nsm.buffalo.edu

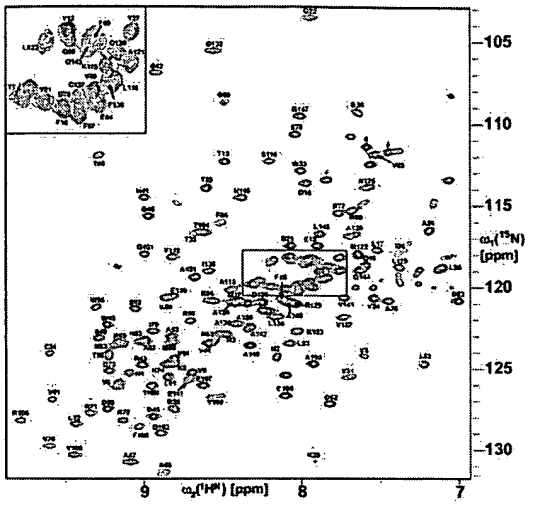


Figure 1. 2D [^{1}H,^{15}N]-HSQC spectrum of CC1736 from *C. crescentus*. Peaks are labeled with their respective sequential resonance assignments using the one-letter code of amino acids and the amino acid sequence number. For clarity, the assignments of the crowded central region of the spectrum are shown in an insert at the upper left corner.

spectrometer operating at 750 MHz ^{1}H resonance frequency, processed with the program PROSA (Güntert et al., 1992) and analyzed using the program XEASY (Bartels et al., 1995). Sequence specific backbone (H^N, H^α, N, C′, C^α) and H^β/C^β resonance assignments were achieved by using the reduced dimensionality (RD) experiments 3D $\underline{H}^{\alpha/\beta}\underline{C}^{\alpha/\beta}$(CO)NHN and HNN<u>CAHA</u> (Szyperski et al., 1998; 2002) in conjunction with conventional 3D HNNCACB and 3D HNN(CO)CA (Cavanagh et al., 1996). Side chain ^{1}H and ^{13}C resonance assignments were obtained using RD 3D <u>HCCH</u>-COSY (Szyperski et al., 2002). 3D ^{13}C- and ^{15}N-resolved [^{1}H-^{1}H] NOESY spectra (Cavanagh et al., 1996) were recorded to support (i) the sequential resonance assignment and (ii) the identification of regular secondary structure elements by observation of ^{1}H–^{1}H nuclear Overhauser enhancements (NOEs).

Extent of assignments and data deposition

Nearly complete sequential resonance assignment were obtained for CC1736, that is, 95.9% of the backbone shifts, excluding the N-terminal NH_3^+, the Pro ^{15}N and the ^{13}C′ shifts of residues preceding the Pro residues, and $^{13}C^\beta$ resonances, and 94.7% of the side chain chemical shifts, excluding the Lys NH_3+, the Arg NH_2, the OH, the side chain ^{13}C′ and the aromatic quaternary ^{13}C shifts, were obtained. The assigned 2D [^{15}N, ^{1}H] HSQC spectrum is shown in Figure 1. Chemical shifts are deposited in the BioMagResBank (accession number 6120). Based on these chemical shifts, and supported by observation of sequential and medium-range NOEs, the regular secondary structure elements of CC1736 were identified. Seven β-strands (residues 3–10, 34–44, 47–56, 62–71, 76–82, 89–98, 101–111) and two α-helices (residues 15–21, 116–144) were found, demonstrating that CC1736 belongs to the class of α/β proteins. The ongoing 3D structure determination will provide the first 3D structural information for a protein domain family of currently unknown function. Moreover, the CC1736 structure will be among the first to be made available for *C. crescentus*.

Acknowledgements

This work was supported by the Protein Structure Initiative of the *National Institutes of Health* (P50 GM62413) and the *National Science Foundation* (MCB 00075773 to T.S.).

References

Bartels, C., Xia, T., Güntert, P., Billeter, M. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

Cavanagh, J., Fairbrother, W.J., Palmer III, A.G. and Skelton, N.J. (1996) *Protein NMR Spectroscopy*, Wiley, New York, NY.

Güntert, P., Dötsch, V., Wider, G. and Wüthrich K. (1992). *J. Biomol. NMR*, **2**, 619–629.

Jansson, M., Li, Y.-C., Jendeberg, L., Anderson, S., Montelione, G. T. and Nilsson, B. (1996) *J. Biomol. NMR*, **7**, 131–141.

Laub, M.T., McAdams, H., Fledblyum, T., Fraser, C. and Shapiro, L. (2000) *Science*, **290**, 2144–2148.

Nierman, W.C. et al. (2001) *Proc. Natl. Acad. Sci. USA*, **98**, 4136–4141.

Szyperski, T., Banecki, B. Braun, D. and Glaser, R. (1998) *J. Biomol. NMR*, **11**, 387–405.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N. and Montelione, G.T. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 8009–8014.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wunderlich, Z., Acton, T.B., Liu, J., Kornhaber, G., Everett, J., Carter, P., Lan, N., Echols, N., Gerstein, M., Rost, B. and Montelione, G.T. (2004) *Protein Sci.*, in press.

EXHIBIT 21

PROTEINS: Structure, Function, and Bioinformatics 54:794–796 (2004)

STRUCTURE NOTE

NMR Structure of the Hypothetical Protein AQ-1857 Encoded by the Y157 Gene From *Aquifex aeolicus* Reveals a Novel Protein Fold

Duanxiang Xu,[1,4] Gaohua Liu,[1,4] Rong Xiao,[2,4] Tom Acton,[2,4] Sharon Goldsmith-Fischman,[3,4] Barry Honig,[3,4] Gaetano T. Montelione,[2,4] and Thomas Szyperski[1,4]*

[1]*Department of Chemistry, University at Buffalo, The State University of New York, Buffalo, New York*
[2]*Center of Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey*
[3]*Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[4]*Northeast Structural Genomics Consortium*

Introduction. The open reading frame of the hypothetical protein AQ-1857 (SwissProt/TrEMBL ID Y157_AQUAE) encoded in the genome of *Aquifex aeolicus* was selected[1] as a target for the Northeast Structural Genomics Consortium (NESGC; http://www.nesg.org). Here we report the high-quality NMR solution structure of AQ-1857 (NESG target QR6). The 116-residue protein AQ-1857 belongs to the HesB/YADR/YFHF family[2,3] which includes proteins involved in nitrogen fixation[4] and Fe-S cluster assembly,[5,6] and contains the PROSITE[7] consensus pattern for this family:

F-X-[LIVMFY]-X-N-[PG]-[NSKQ]-
X(4)-C-X-C-[GS]-X-S-F.

Materials and Methods. Uniformly (*U*) ^{13}C, ^{15}N-labeled AQ-1857 was cloned, expressed and purified following standard protocols. Briefly, the full length gene (Y157_AQUAE) from *Aquifex aeolicus* was cloned into a pET21d (Novagen) derivative, yielding the plasmid pQR6-21. The resulting construct contains eight nonnative residues at the C-terminus (LEHHHHHH) that facilitate protein purification. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with pQR6-21, and cultured in MJ minimal medium containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as sole nitrogen and carbon sources. *U*-^{13}C,^{15}N AQ-1857 was purified using a two-step protocol consisting of Ni-NTA affinity (Qiagen) and gel filtration (HiLoad 26/60 Superdex 75, Amersham Biosciences) chromatography. The final yield of purified *U*-^{13}C, ^{15}N AQ-1857 (> 97% homogeneous by SDS-PAGE; 14.4 kDa by MALDI-TOF mass spectrometry) was about 10 mg/L. In addition, a sample which was *U*-^{15}N and 5% biosynthetically directed fractionally ^{13}C-labeled was generated for stereospecific assignment of isopropyl methyl groups.[8] Two samples of 5%^{13}C,*U*-^{15}N and *U*-^{13}C,^{15}N AQ-1857 were prepared at concentrations of 1.0 mM in 95% H_2O/5% D_2O solution containing 20 mM MES, 100 mM NaCl, 10 mM DTT, 5 mM $CaCl_2$, 0.02% NaN_3 at pH 6.5.

All NMR data were collected at 20°C on Varian INOVA 600 and 750 spectrometers. The spectra were processed and analyzed using the programs NMRPipe[9] and XEASY,[10] respectively. Resonance assignments were obtained as described[11] using a suite of reduced-dimensionality NMR experiments, including 3D HNNCAHA, $H^{\alpha\beta}C^{\alpha\beta}$(CO)NHN, HCCH-COSY, and 2D HBCB(CGCD)HD. These data were complemented by conventional[12] HNNCACB and HC(C)H TOCSY experiments. Assignments were obtained for 93% of the backbone and $^{13}C^{\beta}$, and for 91% of the side chain chemical shifts. Stereospecific assignments were obtained for 44% of the β-methylene groups exhibiting non-degenerate proton chemical shifts, and for all Val and Leu isopropyl moieties. The chemical shifts were deposited in the BioMagResBank (accession code: 5683). Upper distance limit constraints for structure calculations were obtained from 3D ^{15}N-and ^{13}C-resolved [^1H,^1H]-NOESY[12] (Table I). In addition, $^3J_{HN\alpha}$ scalar couplings measured in 3D HNNHA[12] yielded φ-angle constraints, and backbone dihedral angle constraints were derived from chemical shifts as described[13] for residues located in regular secondary structure elements (Table I). Structure calculations were performed using the program DYANA.[14]

Grant sponsor: National Institutes of Health; Grant number: P50 GM2413-10; Grant sponsor: the National Science Foundation; Grant numbers: MCB 00075773, DBI-9904841; Grant sponsor: the Center for Computational Research.

*Correspondence to: Thomas Szyperski, Department of Chemistry, University at Buffalo, The State University at New York, Buffalo, New York 14260. E-mail: szypersk@chem.buffalo.edu

Received 18 August 2003; Accepted 21 August 2003

Results and Discussion. Statistics for the structure determination (Table I) show that a high-quality NMR structure was obtained (Fig. 1). AQ-1857 (PDB ID: 1NWB) contains seven β-strands A to F and two α-helices. A(↓), F(↓) and G(↑) form a 3-stranded, and D(↓), E(↑), B(↑) and C(↓) form a 4-stranded sheet. The two sheets form a "sandwich" being rotated by ~45 degrees relative to each other (Fig. 2). The segment 40–45 and the C-terminal tail 102–116 are flexibly disordered in solution.

The NMR structure of AQ-1857 is the first structure representative of the larger HesB family[2,3] of proteins. No meaningful structural homologues were identified using the programs SKAN,[15] DALI,[16] or CE.[17] This finding strongly supports the notion that AQ-1857 possesses a

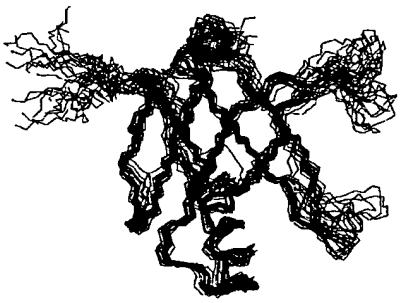


Fig. 1. The 20 DYANA conformers with the lowest residual DYANA target function chosen to represent the NMR solution structure of AQ-1857 are shown after superposition of the backbone heavy atoms N, C^{α} and C′ of the regular secondary structure elements for minimal RMSD.

TABLE I. Statistics of 20 Best DYANA Conformers of AQ-1857†

Distance constraints	
All	1328
Intraresidue [$i = j$]	274
Sequential [$(i - j) = 1$]	428
Medium Range [$1 < (i - j) \leq 5$]	242
Long Range [$(i - j) > 5$]	384
Dihedral angle constraints	
ϕ	74
ψ	62
Number of constraints per residue	14.4
Number of long-range constraints per residue	3.8
Average pairwise RMSD (Å) to the mean coordinates	
All residues[a]	
Backbone atoms	1.24 ± 0.24
All heavy atoms	1.77 ± 0.22
Regular secondary structure elements[b]	
Backbone atoms	0.59 ± 0.17
All heavy atoms	1.06 ± 0.16
Distance constraints violations per conformer	
0.1–0.2 Å	1.75
0.2–0.5 Å	1.0
>0.5 Å	0
Dihedral-angle constraint violation per conformer	
0–10°	0.15
>10°	0
Ramachandran Plot	
Residues in most favored regions (%)	81
Residues in additional allowed regions (%)	18
Residues in generously allowed regions (%)	1
Residues in disallowed regions (%)	0

†20 conformers with lowest DYANA target function values (0.78 ± 0.16 $Å^2$; range: 0.45–1.04 $Å^2$) out of 100 calculated.
[a]Residues 1–101; the C-terminal segment 102–116 is flexibly disordered in solution.
[b]Residues 14–25 and 77–79 (α-helices), and 10–12, 33–36, 52–53, 63–65, 69–72, 84–89, 94–99 (β-strands).

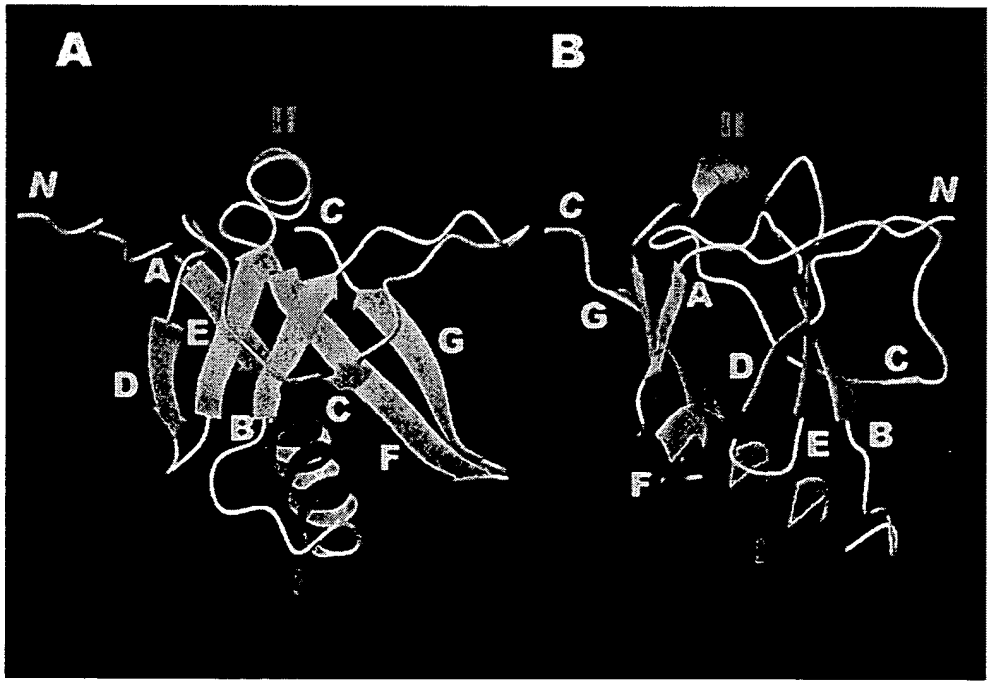


Fig. 2. **A:** The novel fold of AQ-1857: ribbon drawing of the DYANA conformer with the lowest residual target function value. The α-helices I and II are shown in red and yellow, the β-strands A to G are in cyan, other polypeptide segments are in grey, and the N- and C-terminal ends of the protein are indicated as 'N' and 'C'. **B:** Same as in (A), but rotated by 90° about the vertical axis. α-Helix I: residues 14–25, II: 77–79; β-Strand A: 10–12, B: 33–36, C: 52–53, D: 63–65, E: 69–72, F: 84–89, G: 94–99.

hitherto uncharacterized, novel fold (Fig. 2). Interestingly, the PROSITE consensus pattern for the HesB family spans the flexibly disordered C-terminal tail of the protein. In fact, two of the three cysteinyl residues which have been proposed to be involved in iron-sulfur cluster assembly in members of this family[5,6] are located in this tail (not shown in Fig. 1; the third cysteine is located in position 43 in the loop connecting β-strands 2 and 3). It is thus very likely that the flexibly disordered tail is of functional importance, and it is tempting to speculate that this tail adopts an ordered conformation only upon involvement in Fe-S cluster assembly.

Acknowledgements. This work was supported by the National Institutes of Health (P50 GM62413-01), the National Science Foundation (MCB 00075773 to T.S.; DBI-9904841 to B.H), and the *Center for Computational Research* at UB.

REFERENCES

1. Liu, J, Rost, B. Target space for structural genomics revisited. Bioinfomatics 2002;18:922–933.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucl Acids Res 2002;30:276–280.
3. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucl Acids Res 2003;31:315–318.
4. Huang TC, Lin RF, Chu MK, Chen HM. Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium *Synechococcus sp.* Strain RF-1. Microbiology 1999;145:743–753.
5. Ollagnier-de Choudens S, Nachin L, Sanakis Y, Loiseau L, Barras F, Fontecave M. SufA from *Erwinia chrysanthemi.* Characterization of a scaffold protein required for iron-sulfur cluster assembly. J Biol Chem 2003;278:17993–18001.
6. Wollenberg M, Berndt C, Bill E, Schwenn JD, Seidler A. A dimer of the FeS cluster biosynthesis protein IscA from cyanobacteria binds a [2Fe2S] cluster between to protomers and transfers it to [2Fe2S] and [4Fe4S] apo proteins. Eur J Biochem 2003;279:1662–1671.
7. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bionform 2002;3: 265–274.
8. Szyperski, T, Neri, D, Leiting, B, Otting, G, Wüthrich, K. Support of ^{1}H NMR assignments in proteins by biosynthetically directed fractional ^{13}C-labeling. J Biomol NMR 1992;2:323–334.
9. Delaglio, F., Grzesiek, S., Vuister, GW, Zhu, G, Pfeifer, J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6:277–293.
10. Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 1995;6:1–10.
11. Szyperski, T, Yeh, DC, Sukumaran, DK, Moseley, HNB, Montelione, GT. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 2002;99:8009–8014.
12. Cavanagh J, Fairbrother WJ, Palmer AG, Skelton NJ. Protein NMR spectroscopy. New York: Wiley; 1996.
13. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shifts and sequence homology. J Biomol NMR 1999;13:289–302.
14. Güntert, P, Mummenthaler C, Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
15. Petrey D, Nicholls A, and Honig B. GRASP2: Visualization, surface properties and electrostatics of macromolecular structures. Meth Enzymol 2003;374:494–511.
16. Holm, L, Sander C, Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
17. Shindyalov, IN, Bourne, PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.

EXHIBIT 22

PROTEINS: Structure, Function, and Bioinformatics 55:756–758 (2004)

STRUCTURE NOTE

NMR Structure of the Hypothetical protein NMA1147 From *Neisseria meningitidis* Reveals a Distinct 5-Helix Bundle

Gaohua Liu,[1,4] Dinesh K. Sukumaran,[1] Duanxiang Xu,[1,4] Yiwen Chiang,[2,4] Thomas Acton,[2,4] Sharon Goldsmith-Fischman,[3,4] Barry Honig,[3,4] Gaetano T. Montelione,[2,4] and Thomas Szyperski[1,4]*

[1]*Department of Chemistry, University at Buffalo, the State University of New York, Buffalo, New York*

[2]*Center of Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey*

[3]*Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

[4]*Northeast Structural Genomics Consortium*

Introduction. The 82-residue hypothetical protein NMA1147 (SwissProt/TrEMBL ID Q9JR91) encoded in the genome of *Neisseria meningitidis* is a member of a target sequence cluster[1] of the Northeast Structural Genomics Consortium (NESGC; http://www.nesg.org; NESGC cluster ID 15365) that comprises both eukaryotic and bacterial members. The high-quality NMR solution structure of NMA1147 (NESG target ID MR19) reveals a complex bundle of five α-helices, which is composed of an "up-down" 3-helix and an "orthogonal" 2-helix bundle. NMA1147 belongs to the TPR_div1 Pfam[2] family (Pf03937; DUF339). This family represents a subfamily of the TPR[3] (tetratricopetide repeat) family and exhibits a divergent TPR. The TPR represents an ancient and highly conserved sequence motif spanning two antiparallel helices in proteins involved in various protein-protein interactions.[3] The presence of a variation of this motif in NMA1147 suggests that NMA1147 may be involved in distinct protein-protein interactions, and the location of conserved surface residues identifies a putative interaction surface.

Materials and Methods. Uniformly (*U*) $^{13}C,^{15}N$-labeled NMA1147 was cloned, expressed and purified following standard protocols. Briefly, the full length gene (NMA1147) from *Neisseria meningitidis* was cloned into a pET21 (Novagen) derivative, yielding the plasmid MR19-21. The resulting construct contains eight nonnative residues at the C-terminus (LEHHHHHH) that facilitate protein purification. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with MR19-21, and cultured in MJ minimal medium containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as sole nitrogen and carbon sources. *U*-$^{13}C,^{15}N$ NMA1147 was purified using a two-step protocol consisting of Ni-NTA affinity (Qiagen) and gel filtration (HiLoad 26/60 Superdex 75, Amersham Biosciences) chromatography. The final yield of purified *U*-$^{13}C,^{15}N$ NMA1147 (> 97% homogeneous by SDS-PAGE; 11.4 kDa by MALDI-TOF mass spectrometry) was about 50 mg/L. In addition, a sample which was *U*-^{15}N and 5% biosynthetically directed fractionally ^{13}C-labeled was generated for stereospecific assignment of isopropyl methyl groups.[4] Two samples of 5% ^{13}C,*U*-^{15}N and *U*-$^{13}C,^{15}N$ NMA1147 were prepared at concentrations of 1.0 mM in 95% H_2O/5% D_2O solution containing 20 mM MES, 100 mM NaCl, 10 mM DTT, 5 mM $CaCl_2$, 0.02% NaN_3 at pH 6.5.

NMR data were collected at 25°C on Varian INOVA 600 and 750 spectrometers, and spectra were processed and analyzed using the programs NMRPipe[5] and XEASY.[6] Resonance assignments were obtained as described[7] using a suite of reduced-dimensionality (RD) NMR experiments, including 3D HACA(CO)NHN, HNNCAHA, $H^{\alpha\beta}C^{\alpha\beta}$(CO)NHN, and 2D HBCB(CGCD)HD and ^{1}H-TOCSY relayed HCH-COSY. These data were complemented with conventional[8] HNNCACB and HC(C)H TOCSY experiments. Assignments were obtained for 98% of the backbone and $^{13}C,\beta$ and for 96% of the side chain chemical shifts. Stereospecific assignments were obtained for 18% of the β-methylene groups exhibiting non-degenerate proton chemical shifts, as well as for all Val and Leu isopropyl moieties. The chemical shifts were deposited in the BioMagResBank (Accession code: 5846). Upper distance limit constraints for structure calculations were obtained from 3D ^{15}N- and ^{13}C-resolved [^{1}H,^{1}H]-NOESY[8] (Table I). In addition, $^3J_{HN\alpha}$ scalar couplings measured in 3D HNNHA[8] yielded φ-angle constraints, and backbone dihedral angle constraints were derived from chemical shifts as described[9] for α-helical

Grant sponsor: the National Institutes of Health; Grant number: P50 GM62413-01; Grant sponsor: the National Science Foundation; Grant numbers: MCB 00075773, DBI-9904841; Grant sponsor: the Center for Computational Research

*Correspondence to: Thomas Szyperski, Department of Chemistry, University at Buffalo, The State University at New York, Buffalo, New York 14260. E-mail: szypersk@chem.buffalo.edu

Received 5 September 2003; Accepted 5 September 2003

Published online 14 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20009

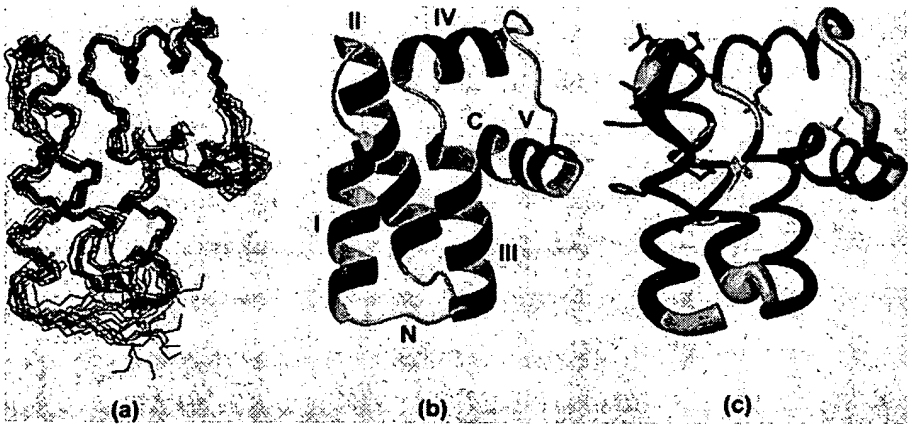


Fig. 1. **a:** The 20 DYANA conformers with the lowest residual DYANA target function chosen to represent the NMR solution structure of NMA1147 are shown after superposition of the backbone heavy atoms N, C^{α} and C′ of the five α-helices I to V for minimal RMSD Helices I and II form an "up-down" bundle, while helices III to IV are arranged in an "orthogonal" manner. **b:** Ribbon drawing of the DYANA conformer with the lowest residual target function value. The α-helices are shown in red and yellow, other polypeptide segments are in grey, and the N- and C-terminal ends of the protein are indicated as "N" and "C." **c:** Molecule core of the backbone of NMA1147 in the standard orientation of (a). For the presentation of the backbone a spline function was draw through the C^{α} positions; the thickness of the cylindrical rod is proportional to the mean of the global displacements of the 20 DYANA conformers calculated after superposition as in described for (a). The helices are shown in red, other polypeptide segments are displayed in grey, the conserved side chains of the molecular core are shown in yellow, and the side chains of conserved surface residues are depicted in blue. For clarity, the backbone of the first three residues, which are flexibly disordered in solution [Fig. 1(a)], is not shown.

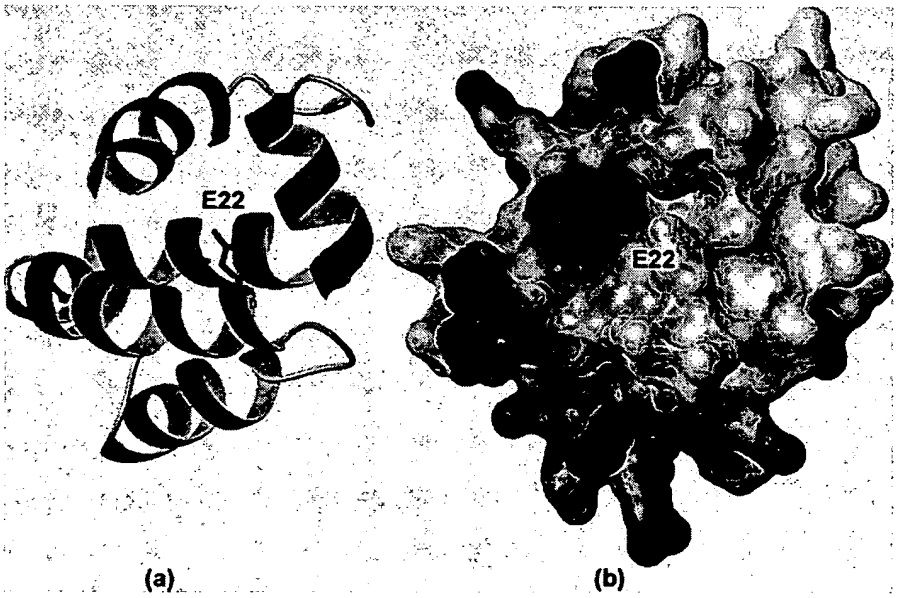


Fig. 2. **a:** Ribbon drawing of NMA1147 [see Fig. 1(b)] in an orientation in which the conserved residue Glu 22 (side chain depicted in magenta) points toward the observer. **b:** Electrostatic surface potential of NMR1147 in the orientation of (a), demonstrating that a negatively charged pocket is located at about the center of a larger hydrophobic patch. Positive and negative charge densities are colored red and blue, respectively.

residues (Table I). Structure calculations were performed using the program DYANA.[10]

Results and Discussion. The statistical parameters summarized in Table I are indicative of a high-quality NMR structure [Fig. 1(a)]. NMA1147 (PDB ID: 1PUZ) is composed of five α-helices I to V [Fig. 1(b)]. The program CATH[11] assigns the NMA1147 structure to the class "mainly α-helical" and to the architectures "up-down" or "orthogonal" bundle. In fact, helices I, II, and III form an up-down 3-helix bundle, while helices IV and V are arranged in an orthogonal manner. The juxtaposition of these two bundles yields a complex 5-helix bundle [Fig.

TABLE I. Statistics of 20 Best DYANA Conformers of NMA1147[a]

Distance constraints	
All	1287
Intraresidue [i = j]	419
Sequential [(i − j) = 1]	376
Medium range [1 < (i − j) ≤ 5]	323
Long range [(i − j) > 5]	169
Dihedral angle constraints	
ϕ	63
Ψ	60
Number of constraints per residue	17.2
Number of long-range constraints per residue	2.1
Average pairwise RMSD (Å) to the mean coordinates	
All residues[b]	
Backbone atoms	0.94 ± 0.26
All heavy atoms	1.49 ± 0.21
Regular secondary structure elements[c]	
Backbone atoms	0.57 ± 0.14
All heavy atoms	1.07 ± 0.14
Distance constraints violations per conformer	
0.2–0.5 Å	0.75
> 0.5 Å	0
Dihedral-angle constraint violation per conformer	
> 5°	0
Ramachandran plot	
Residues in most favored regions (%)	83
Residues in additional allowed regions (%)	16
Residues in generously allowed regions (%)	1
Residues in disallowed regions (%)	1

[a]20 conformers with lowest DYANA target function values (0.99 ± 0.13 Å^2; range: 0.73–1.20 Å^2) out of 100 calculated.
[b]Residues 1–82 (excludes the C-terminal HIS-tag).
[c]Residues, 6–12, 21–36, 40–51, 54–62, 71–82 (α-helices).

1(b)]. Residues conserved among NMA1147 homologues [Fig. 1(c)] are located either in the molecular core, or form a surface cluster involving the last two turns of helix I (Lys 11, Phe 14, Gln 15, Arg 17), the segment connecting helices I and II (Arg 18 and Gly 19), and the N-terminal ends of helices II (Leu 21, Glu 22, Asp 24) and IV (Glu 56). Remarkably, Asp 54, which is neither surface-accessible nor forming a salt bridge in the molecular core, is likewise conserved. Electrostatic surface calculations reveal the presence of an acidic pocket in the region of the conserved Glu 22 at the center of a larger hydrophobic patch (Fig. 2). This pocket is also proximal to conserved surface residues, suggesting that this site is of functional importance.

NMA1147 belongs to the TPR_div1 family (Pf03937; DUF339), a subfamily containing TPR[3]-like motifs. The consensus sequence motif of the TPR[3] repeat, X3-[W/L/Y]-X2-[L/I/M]-[G/A/S]-X2-[Y/L/F]-X8-[A/S/E]-X3-[P/Y/L]-X2-[A/S/L]-X4-[P/K/E], is divergent in NMA1147. The TPR-like sequence spans helices II and III (residues 21–53), in which, however, Phe 27 and Phe 46 are located at positions of relatively small amino acids ([G/A/S] and [A/S/L], respectively) present in the TPR motif. Moreover, the packing of the consensus residues in NMA1147 differs from that found in the TPR motif. Notably, TPR motifs are often occurring as multiple repeats[3] whereas NMA1147 contains only a single motif.

The programs SKAN,[12] DALI,[13] and CE[14] identify structural homologues which are classified by the program SCOP[15] as having either a cyclin-like or a SAM (Sterile Alpha Motif) domain-like fold.[2] Examples include cyclin (1VIN, 1BU2), the TFIIB core domain (1TFB, 1AIS), Recombinase XerD (1A0P) or the RuvA middle domain (1BVS). The structural alignments between NMA1147 and these domains reveal that functional residues in the structural homologues do not align with residues of similar type in NMA1147, nor do they align with the conserved residues in the TPR family [Fig. 1(c)]. Taken together, our analyses thus suggest that the fold of NMA1147 may represent a new SCOP family, or, considering the low sequence identity detected between NMA1147 and its structural homologues, a new SCOP *super*family. Members of this new (super)family may possibly function to mediate protein-protein interactions in a manner that is distinct from that of the classical TPR motif.

Acknowledgments. This work was supported by the National Institutes of Health (P50 GM62413-01), the National Science Foundation (MCB 00075773 to T.S., DBI-9904841 to B.H), and the Center for Computational Research at UB.

REFERENCES

1. Liu, J, Rost, B. Target space for structural genomics revisited. Bioinfomatics 2002;18:922–933.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
3. Blatch GL, Lassle M. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions, BioEssays 1999;21: 932–939.
4. Szyperski T, Neri D, Leiting B, Otting G, Wüthrich K. Support of ^{1}H NMR assignments in proteins by biosynthetically directed fractional ^{13}C-labeling. J Biomol NMR 1992;2:323–334.
5. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6:277–293.
6. Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 1995;6:1–10.
7. Szyperski T, Yeh DC, Sukumaran DK, Moseley HNB, Montelione GT. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 2002;99:8009–8014.
8. Cavanagh J, Fairbrother WJ, Palmer AG, Skelton NJ. Protein NMR spectroscopy. New York: Wiley; 1996.
9. Cornilescu, G, Delaglio, F, Bax, A. Protein backbone angle restraints from searching a database for chemical shifts and sequence homology. J Biomol NMR 1999;13:289–302.
10. Güntert, P, Mummenthaler C, Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
11. Orengo CA, Michie AD, Jones S, Jones DT, Swindles MB, and Thornton, JM. CATH–A hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
12. Petrey D, Nicholls A, Honig B. GRASP2: Visualization, surface properties and electrostatics of macromolecular structures and sequences. Meth Enzymol 2003;374:492–509.
13. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
14. Shindyalov IN, Bourne, PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
15. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.

EXHIBIT 23

PROTEINS: Structure, Function, and Bioinformatics 58:747–750 (2005)

STRUCTURE NOTE

NMR Structure of the 18 kDa Protein CC1736 From *Caulobacter crescentus* Identifies a Member of the "START" Domain Superfamily and Suggests Residues Mediating Substrate Specificity

Yang Shen,[1,4] Sharon Goldsmith-Fischman,[2,4] Hanudatta S. Atreya,[1,4] Thomas Acton,[3,4] LiChung Ma,[3,4] Rong Xiao,[3,4] Barry Honig,[2,4] Gaetano T. Montelione,[3,4] and Thomas Szyperski[1,4]*

[1]*Department of Chemistry, University at Buffalo, the State University of New York, Buffalo, New York*
[2]*Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[3]*Center of Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey*
[4]*Northeast Structural Genomics Consortium*

Introduction. The 18-kDa protein CC1736 (SwissProt/TrEMBL ID Q9A7I7) of *Caulobacter crescentus* belongs to a target cluster[1] of the Northeast Structural Genomics Consortium (NESGC; http://www.nesg.org; NESGC target ID: CcR19; NESGC Rost-cluster ID: 17538) comprising both eukaryotic and bacterial proteins. CC1736 is member of Pfam[2] "Aromatic-rich protein family" (Pf03654) and has at least 90 close sequence homologs. CC1736 exhibits weak sequence homology with oligoketide cyclases and aromatases which bind multi-cyclic and/or aromatic compounds such as cholesterol (or polyketides).[3]

Materials and Methods. As described elsewhere,[4] resonance assignments were obtained from NMR data collected for a $^{13}C,^{15}N$-labeled CC1736 sample (1.1 mM) at 25°C using a Varian INOVA 750 spectrometer. Upper-distance limit constraints were obtained from 3D ^{15}N- and ^{13}C-resolved [$^1H,^1H$]-NOESY[5] (Table I), $^3J_{HN\alpha}$ scalar couplings were measured in 3D HNNHA[5] yielding ϕ-angle constraints, and backbone dihedral angle constraints were derived from chemical shifts as described[6] for residues in secondary structure elements. Structure calculations were performed using the program DYANA.[7]

Results and Discussion. The NMR structure [Table I, Fig. 1(a)] of CC1736 (PDB ID: 1T17) reveals two α-helices I and II and seven β-strands A–F [Fig. 1(b)] giving rise to an $\alpha+\beta$ fold in which helices and strands are segregated. The β-strands are arranged with topology A($\uparrow$), G($\downarrow$), F($\uparrow$), E($\downarrow$), D($\uparrow$), C($\downarrow$), B($\uparrow$) and form a highly twisted β-sheet. The juxtaposition of α-helices and β-strands leads to the formation of a hydrophobic tunnel, which is likely of functional importance (see below).

A search for structurally similar domains using the programs DALI,[8] SKAN,[9] and CE[10] reveals that CC1736 is structurally similar to birch pollen allergen (PDB ID: 1BV1, DALI Z-score 12.0), phosphatidylinositol transfer protein (PDB ID: 1FVZ, DALI Z-score 8.1), phosphoglucomutase domain (PDB ID: 3PMG, DALI Z-score 6.0), and the cholesterol-regulated START domain (PDB ID: 1JSS, DALI Z-score 11.5). According to SCOP[11] classification, these proteins exhibit a "TBP-like" (TATA-Binding Protein-like) fold and they belong to the START Domain Superfamily,[12] which includes oligoketide cyclases.

Sequence comparisons show that CC1736 is: i) a member of the "Aromatic-Rich Protein Family" (Pfam[2] ID: PF03654), which is annotated as having a possible relationship to polyketide synthases/cyclases, and ii) exhibits modest sequence homology with oligoketide cyclases such as the WhiE putative polyketide cyclase (26% identity, BLAST E-value = 0.17, member of "Polyketide Cyclase Family" Pfam ID: PF03364). No sequence similarity is detected between CC1736 and START domain even in iterative PSI-BLAST[13] searches. However, the fold recognition program HMAP[14] recognizes a relationship between CC1736 and START domain,[8] supporting without reference to its NMR structure that CC1736 is a member of the START domain superfamily.

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: the National Institutes of Health; Grant number: P50 GM62413-01; Grant sponsor: the National Science Foundation; Grant numbers: MCB 00075773, DBI 9904841; Supported by: UB Center for Computational Research.

*Correspondence to: Thomas Szyperski, Department of Chemistry, University at Buffalo, The State University at New York, Buffalo, New York 14260. E-mail: szypersk@chem.buffalo.edu

Received 31 August 2004; Accepted 30 September 2004

Published online 22 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20365

TABLE I. Statistics of 20 Best DYANA[7] Conformers of CC1736

Conformationally restricting distance constraints	
Intraresidue [i = j]	568
Sequential [(i − j) = 1]	789
Medium Range [1 < (i − j) ≤ 5]	332
Long Range [(i − j) > 5]	579
Total	2268
Dihedral angle constraints	
ϕ	136
ψ	106
Number of constraints per residue	17.0
Number of long-range constraints per residue	3.9
Completeness of stereospecific assignments[a] [%]	
$^{\alpha}CH_2$ of Gly	88 (7/8)
$^{\beta}CH_2$	51 (34/67)
Val and Leu isopropyl groups	65 (13/23)
DYANA target function [Å^2]	1.38 ± 0.17
Average RMSD to the mean DYANA coordinates [Å]	
Regular secondary structure elements[b], backbone heavy atoms N, C$^{\alpha}$, C′	0.44 ± 0.07
Regular secondary structure elements, all heavy atoms	0.92 ± 0.06
Residues 1–148, backbone heavy atoms N, C$^{\alpha}$, C′	0.62 ± 0.09
Residues 1–148, all heavy atoms	1.13 ± 0.09
Heavy atoms of molecular core (or best-defined SC)[c]	0.47 ± 0.06
Ramachandran plot summary for residues 1–148 [%]	
Most favored regions	83
Additionally allowed regions	15
Generously allowed regions	2
Disallowed regions	0
Average number of distance constraints violations per DYANA conformer [Å]	
0.2 − 0.5	1.1
> 0.5	0
Average number of dihedral-angle constraint violations per DYANA conformer [degrees]	
> 5	0

[a]Relative to pairs with nondegenerate chemical shifts.
[b]Residues, 3–10, 37–45, 48–57, 64–73, 78–83, 91–99, 104–112 (β-strands), and 15–21, 117–146 (α-helices).
[c]Includes 49 residues: 4, 6–7, 10–11, 13, 21, 24, 32, 34–35, 39–40, 44, 47–48, 50–51, 53, 55, 57, 67–68, 70, 78–82, 103–104, 106, 110, 113–114, 119, 121, 123–127, 130–131, 135–137, 140, 142.

Recently, the X-ray structure[15] of the polyketide cyclase SnoaL (PDB ID 1SJW) was solved. SnoaL (1SJW) and CC1736 exhibit very low sequence identity (~ 5% based on the CE structural alignment) and quantitative comparison of the CC1736 and SnoaL coordinates using the programs DALI (Z-score 2.9; RMSD 3.4 Å for 77 aligned residues) and CE (Z-score 3.5; RMSD 3.8 Å for 88 aligned residues) did not reveal any sigificant structural similarity. However, visual comparison of CC1736 and SnoaL reveals similar architectures, and the binding of the ligand of SnoaL (nogalonic acid methyl ester) in the center of a hydrophobic tunnel suggests that similar bioreactions are possibly catalyzed by the two proteins. The finding that CC1736 belongs to the START domain superfamily allows one to identify residues which likely confer functional specificity. Members of the START domain superfamily[12] are believed to bind polycyclic compounds such as cholesterol and polyketides in a hydrophobic tunnel.[15–17] CC1736 contains several conserved hydrophobic and aromatic residues lining this tunnel (Fig. 1). Structural alignment of CC1736 with the cholesterol-binding START domain[16] (RMSD between CC1736 and the START domain: 2.5 Å for 131 structurally aligned residues) indicates that residues proposed to mediate the specificity of the START domain for cholesterol correspond in CC1736 to residues having different physical–chemical properties. For example, Ala 55 and Val 70 of CC1736 correspond to a buried charged pair in the START domain[16] which is proposed to convey specificity for cholesterol binding.[18] This suggests that CC1736 and START domain maintain different substrate specificities. Structural analysis [Fig. 1(d and e)] reveals that the tunnel in CC1736 exhibits a positive electrostatic potential at one end (due to Lys 8) and contains one polar side chain oriented towards the inside of the tunnel (Asn 93). A small cavity affords access to the side chain of Asn 93. In addition, a salt-bridge formed between Glu 64 and Lys 115 is located at the entry to the tunnel. It appears likely that these residues, in conjunction with Ala 55 and Val 70, participate in defining ligand specificity for CC1736 [Fig. 1(f)].

Further insights into functional specificity and evolutionary origin were obtained by investigating the genomic organization surrounding genes of the CC1736 family using the SwissProt[19] and TIGR[20] databases. Two distinct genomic organizations are observed for bacterial species. In *C. crescentus*, gene CC1736 is located proximal to genes encoding components of the pyruvate dehydrogenase complex and LipA. Several bacterial homologs share this "CC1736-type" genomic organization, and eukaryotic genes exhibit the same pattern of conserved residues found in these genomes (Fig. 1S of Supplementary Information). In the second class of bacterial homologs, the gene homologous to CC1736 is located proximal to those encoding an RNA binding protein ("ssRA") and a hypothetical protein belonging to a family of proteins of unknown function (Pfam ID: UPF0125). Intriguingly, a correlation is found between the conservation of residues that are predicted to be involved in substrate specificity based on structure analysis [e.g., Lys 8, Glu 64, Val 70, Asn 93, and Lys 115 in CC1736; Fig. 1(d–f)] and the genomic organization. Bacterial homologs containing the "CC1736-type" genomic organization conserve the residues predicted to be involved in CC1736 substrate specificity, which are also maintained in the eukaryotic homologs. In contrast, in the bacterial homologs with the second type of genomic organization, conserved residues are found at the positions corresponding to the conserved residues of the CC1736-type genomes. However, these residues have quite different physical chemical properties (Fig. 1S of Supplementary Information). This suggests that the CC1736 family of sequence homologs for which the genome organization is currently known, can be divided into two classes, each of which very likely binds ligands of different chemical nature. A phylogenetic tree constructed with the program PhyloDraw (Fig. 2) shows that the classes cluster in distinct parts of the tree. We thus conclude that: i) the CC1736 gene is of

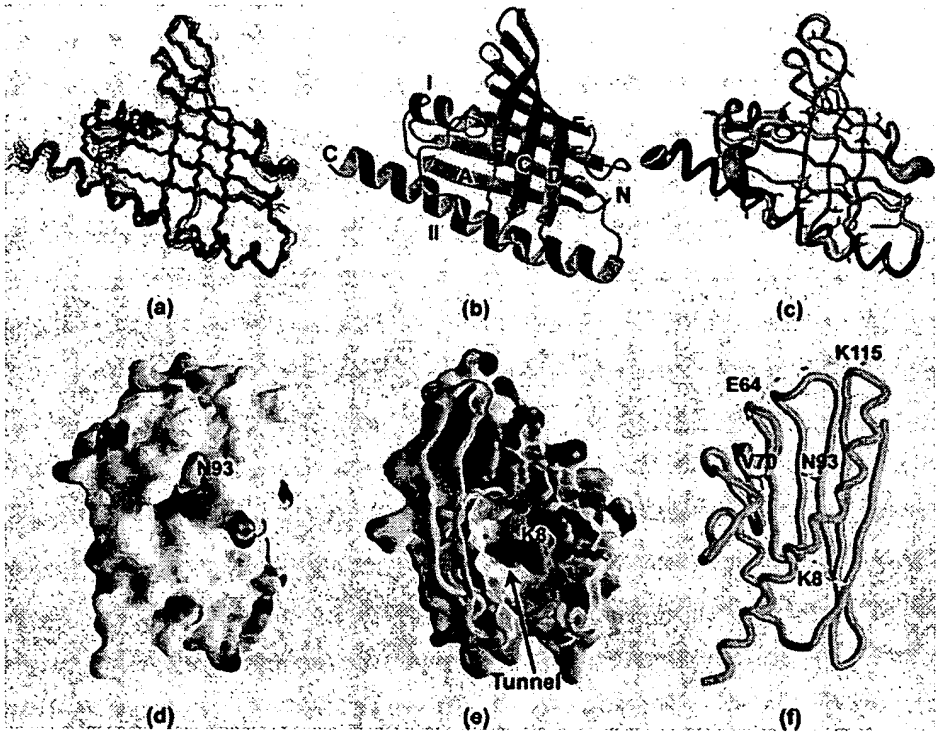


Fig. 1. NMR structure of CC1736. **a:** The 20 DYANA conformers with the lowest residual DYANA target function representing the NMR solution structure of CC1736 are shown after superposition of the backbone heavy atoms N, C^{α} and C′ of the residue 2 to residue 146 for minimal RMSD. **b:** Ribbon drawing of the DYANA conformer with the lowest residual target function value (Table I). The α-helices (I and II) are shown in red and yellow, the β-strands (A to G) are in cyan, other polypeptide segments are in grey, and the N- and C-terminal ends of the protein are indicated as "N" and "C." **c:** Backbone and best-defined side chains of CC1736 in the standard orientation of (a). For the presentation of the backbone a spline function was draw through the C^{α} positions and the thickness of the cylindrical rod is proportional to the mean of the global displacements of the 20 DYANA conformers calculated after superposition as in described for (a). The α-helices are shown in red, β-strands in cyan, other polypeptide segments are displayed in grey, and the 49 best-defined side chains (Table I) are shown in yellow. **d:** Electrostatic surface of CC1736 showing the cavity exposing Asn 93. **e:** CC1736 rotated 90° about the x-axis relative to (d) and slabbed along the z-axis to show the hydrophobic tunnel (see text). Lys 8 contributes to the positive electrostatic potential at one end of the tunnel. **f:** Backbone of CC1736 in the orientation shown in (d). Residues proposed to contribute to substrate specificity (see text) are labeled. Figures (a–c) and (d–e) were generated using the programs MOLMOL[21] and GRASP,[9] respectively.

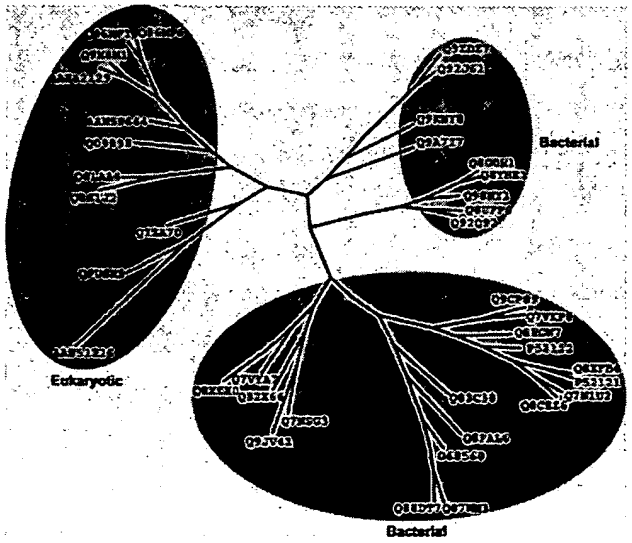


Fig. 2. Phylogenetic tree generated using the program PhyloDraw (http://pearl.cs.pusan.ac.kr/phylodraw) with the sequence alignment of Figure 1 of the Supplementary Information. Genes are represented by their SwissProt ID. Bacterial sequences embedded in a genome with CC1736-type organization (see text) are all located in the red area (Q9A7I7 represents the CC1736 gene), eubacterial sequences with an alternative genome organization (see text) are located in the green area, and eukaryotic sequences are located in the blue area.

ancient evolutionary origin, and that ii) eukaryotic homologs of CC1736 likely function similarly as the bacterial congeners embedded in a CC1736-type genome organization.

Acknowledgments. This work was supported by the National Institutes of Health (P50 GM62413-01), the National Science Foundation (MCB 00075773 to T.S., DBI 9904841 to B.H), and the Center for Computational Research at UB.

REFERENCES

1. Liu, J, Rost, B. Target space for structural genomics revisited. Bioinfomatics 2002;18:922–933.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families data base. Nucleic Acids Res 2002;30:276–280.
3. Shen, B, Hutchinson, CR. Deciphering the mechanism for the assembly of aromatic polyketides by a bacterial polyketide synthase. Proc Natl Acad Sci USA 1996;93:6600–6604.
4. Shen, Y, Atreya, HS, Xiao, R, Acton, TB, Shastry, R, Ma, L, Montelione, GT, Szyperski, T. Resonance assignments for the 18 kDa Protein CC1736 from *Caulobacter crescentus*. J Biomol NMR 2004;29:549–550.
5. Cavanagh, J, Fairbrother, WJ, Palmer, AG, Skelton, NJ. Protein NMR spectroscopy. New York: Wiley; 1996.
6. Cornilescu, G, Delaglio, F, Bax, A. Protein backbone angle re-

straints from searching a database for chemical shifts and sequence homology. J Biomol NMR 1999;13:289–302.
7. Güntert, P, Mummenthaler C, Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
8. Holm, L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
9. Petrey D, Nicholls A, Honig B. GRASP2: Visualization, surface properties and electrostatics of macromolecular structures and sequences. Meth Enzymol 2003;374:492–509.
10. Shindyalov, IN, Bourne, PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
12. Iyer, LM, Koonin, EV, Aravind, L. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. Proteins 2001;43:134–144.
13. Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, Lipman, DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
14. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. On the role of structure and sequence information for remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043–1062.
15. Sultana, A, Kallio P, Jansson, A, Wang, JS, Niemi, J, Mantsala, P, Schneider, G. Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation. EMBO J 2004;23:1911–1921.
16. Romanowski, MJ, Soccio, RE, Breslow, JL, Burley, SK. Crystal structure of the *Mus musculus* cholesterol-regulated START protein 4 (StarD4) containing a StAR-related lipid transfer domain. Proc Natl Acad Sci USA 2002;99:6949–6954.
17. Tsujishita, Y, Hurley, JH. Structure and lipid transport mechanism of a StAR-related domain. Nat Struct Biol 2000;7:408–414.
18. Zhang, M, Liu, P, Dwyer, NK, Christenson, LK, Fujimoto, T, Martinez, F, Comly, M, Hanover, JA, Blanchette-Mackie, EJ, Strauss, JF. MLN64 mediates mobilization of lysosomal cholesterol to steroidogenic mitochondria. J Biol Chem 2002;277:33300–33310.
19. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res 2003;31:365–370.
20. http://www.tigr.org/tdb/
21. Koradi R, Billeter M., Wüthrich K. MOLMOL: A program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–55.

EXHIBIT 24

PROTEIN STRUCTURE REPORT

The NMR solution structure of the 30S ribosomal protein S27e encoded in gene *RS27_ARCFU* of *Archaeoglobus fulgidis* reveals a novel protein fold

CATHERINE HERVE DU PENHOAT,[1,4] HANUDATTA S. ATREYA,[1,4] YANG SHEN,[1,4] GAOHUA LIU,[1,4] THOMAS B. ACTON,[2,4] RONG XIAO,[2,4] ZHAOHUI LI,[3,4] DIANA MURRAY,[3,4] GAETANO T. MONTELIONE,[2,4] AND THOMAS SZYPERSKI[1,4]

[1]Department of Chemistry, University at Buffalo, State University of New York, Buffalo, New York 14260, USA
[2]Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854, USA
[3]Department of Microbiology and Immunology, Weill Medical College of Cornell University, New York, New York 10021, USA
[4]Northeast Structural Genomics Consortium (NEGSC, http://www.nesg.org)

(RECEIVED December 18, 2003; FINAL REVISION January 30, 2004; ACCEPTED February 2, 2004)

Abstract

The *Archaeoglobus fulgidis* gene *RS27_ARCFU* encodes the 30S ribosomal protein S27e. Here, we present the high-quality NMR solution structure of this archaeal protein, which comprises a C4 zinc finger motif of the $CX_2CX_{14-16}CX_2C$ class. S27e was selected as a target of the Northeast Structural Genomics Consortium (target ID: GR2), and its three-dimensional structure is the first representative of a family of more than 116 homologous proteins occurring in eukaryotic and archaeal cells. As a salient feature of its molecular architecture, S27e exhibits a β-sandwich consisting of two three-stranded sheets with topology B(↓), A(↑), F(↓), and C(↑), D(↓), E(↑). Due to the uniqueness of the arrangement of the strands, the resulting fold was found to be novel. Residues that are highly conserved among the S27 proteins allowed identification of a structural motif of putative functional importance; a conserved hydrophobic patch may well play a pivotal role for functioning of S27 proteins, be it in archaeal or eukaryotic cells. The structure of human S27, which possesses a 26-residue amino-terminal extension when compared with the archaeal S27e, was modeled on the basis of two structural templates, S27e for the carboxy-terminal core and the amino-terminal segment of the archaeal ribosomal protein L37Ae for the extension. Remarkably, the electrostatic surface properties of archaeal and human proteins are predicted to be entirely different, pointing at either functional variations among archaeal and eukaryotic S27 proteins, or, assuming that the function remained invariant, to a concerted evolutionary change of the surface potential of proteins interacting with S27.

Keywords: *RS27_ARCFU;* high-throughput NMR; structural genomics; 30S ribosomal protein; zinc finger; *Archaeoglobus fulgidis*

Structural genomics aims at the systematic exploration of protein fold space, with the long-range goal of making the three-dimensional atomic level structure of most proteins easily available form knowledge of the corresponding DNA sequences. In the United States, nine research networks (consortia) are supported through the Protein Structure Initiative set forth by the National Institutes of Health. Among those is the Northeast Structural Genomics Consortium (NEGSC, http://www.nesg.org). In addition to utilizing high-throughput X-ray crystallography, the NESGC is strongly committed to the development of improved NMR (e.g., Montelione et al. 2000; Szyperski et al. 2002; Xia et al. 2002; Yee et al. 2002; Kim and Szyperski 2003; Zheng et al. 2003) and structural bioinformatics methodology

Reprint requests to: Thomas Szyperski, Department of Chemistry, University of Buffalo, State University of New York, 816 Natural Sciences Complex, Buffalo, NY 14260, USA; e-mail: szypersk@chem.buffalo.edu; fax: (716) 645-7338.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03589204.

(Goldsmith-Fischman and Honig 2003) for structural genomics.

Here, we report the NMR solution structure of the protein S27e encoded in gene *RS27_ARCFU* of *Archaeoglobus fulgidis* (Klenk et al. 1997). S27e is a 30S ribosomal protein and was selected as NESGC target GR2, representing a family of more than 116 sequence homologs from eukaryotic and archaebacterial organisms (Fig. 1). Atomic resolution structures have recently been solved for the large subunit of an archaeal ribosome (Ban et al. 2000), as well as the small subunit of a bacterial ribosome (Schluenzen et al. 2000; Brodersen et al. 2002). Most importantly, however, bacterial ribosomes do not contain a S27 homolog. Hence, the S27e structure determination was targeted by the NESGC in order to provide (1) a high-leverage value—that is, a large number of homology models derived from the experimentally determined structure, and (2) novel insights into the operation of the eukaryotic ribosome. Evidently, (1) meets with a primary goal of structural genomics, that is, the exploration of fold space. Because ribosomal proteins operate in the context of the large macromolecular assembly of the ribosome, their function may not be readily inferred from structure alone. Nonetheless, the isolated ribosomal proteins appear to be valuable structural (genomics) targets; many ribosomal proteins retain their RNA-binding specificities and/or have functions outside of the ribosome (Ramakrishnan and White 1998), and the assessment of conformational changes upon ribosome formation promises to lead to new insights into protein–protein and protein–RNA interactions.

Ribosomes represent the central unit of the protein-synthesizing machinery of living cells and consist roughly of two-thirds RNA and one-third protein. The ribosomal proteins are named according to the subunit to which they belong (small subunit: S1 to S31; large subunit: L1 to L44), and they cover a large variety of structural and functional roles (Ramakrishnan and White 1998). In particular, eukaryotic ribosomal protein S27 has been reported to be involved in rRNA processing (BaudinBaillieu et al. 1997), the degradation of damaged mRNAs (Revenkova et al. 1999), as well as direct binding to mRNA (Takahashi et al. 2002). However, the structural basis of S27–RNA interactions remains to be explored. Moreover, the X-ray structures of the large (Ban et al. 2000) and small ribosomal subunits (Schluenzen et al. 2000; Brodersen et al. 2002) reveal that ribosomal proteins may interact with both other proteins and ribonucleic acids. This may eventually be found for S27 proteins.

Several ribosomal proteins contain zinc finger motifs that may serve to mediate protein–RNA interactions (Frankel 2000), and the sequences of S27 genes from various species (Fig. 1) demonstrate that most of them are C4 zinc finger proteins of the $CX_2CX_{14-16}CX_2C$ class. It has been suggested that the S27 zinc finger may be a fossil from ancient evolution (Chan et al. 1993), that is, the zinc finger is possibly not anymore of functional importance for modern S27 proteins. This view is supported by the sequence alignment (Fig. 1) showing that the zinc finger motif is not strictly conserved. Moreover, the eukaryotic proteins have an amino-terminal extension when compared with their archaebacterial congeners. Taken together, it might thus be that eukaryotic and archaeal S27 proteins have evolved divergently in order to undertake different roles in their ribosomes.

Results and Discussion

NMR structure of S27e

A total of 669 conformationally restricting NOE distance constraints were derived from three-dimensional (3D) ^{15}N- and ^{13}C-resolved [^1H,^1H]-NOESY. In addition, 31 $^3J_{HN\alpha}$ coupling constants yielded Φ-angle constraints, and 60 backbone dihedral angle constraints were obtained from chemical-shift values (Cornilescu et al. 1999) for residues located in β-strands. Stereospecific assignments were obtained for one glycine α-methylene proton pair (25% of the pairs with nondegenerate chemical shifts), seven β-methylene proton pairs (23%), 15 more peripheral methylene proton pairs, and for all six valine isopropyl methyl groups. The resulting ensemble of 20 DYANA (Guntert et al. 1997) conformers, together with the corresponding NMR constraints, have been submitted to the PDB (1QXF).

An illustration of the quality of the S27e structure is presented in Figure 2A, which shows the polypeptide backbone of the 20 best DYANA conformers selected to represent the solution structure. The absence of any large constraint violations (Table 1) demonstrates that these experimental constraints are well satisfied in the set of 20 conformers, and the small RMSD values (Table 1) indicates a high-quality NMR structure. Furthermore, plots of local backbone RMSD values and global backbone displacements (Fig. 3) show (1) that the β-sheets are structurally very well defined, and (2) that increased local *and* global disorder is observed for the amino-terminal tetrapeptide segment, the two loops comprising residues 21–25 and 40–45, and the carboxy-terminal segment of residues 53–58. The high quality of the NMR structure is also reflected in the narrow ranges among the 20 DYANA conformers of most Φ and Ψ dihedral angles (Fig. 4).

Three-dimensional structure of S27e

As a salient feature of its molecular architecture, S27e exhibits a β-sandwich consisting of two three-stranded sheets with topology B(↓), A(↑), F(↓), and C(↑), D(↓), and E(↑) (Fig. 2B). Both β-sheets have similar right-handed twists so that in the face-to-face arrangement the strands in the first sheet are almost parallel to those in the second one (Chothia et al. 1977). Notably, strands D and F have bulges at Thr

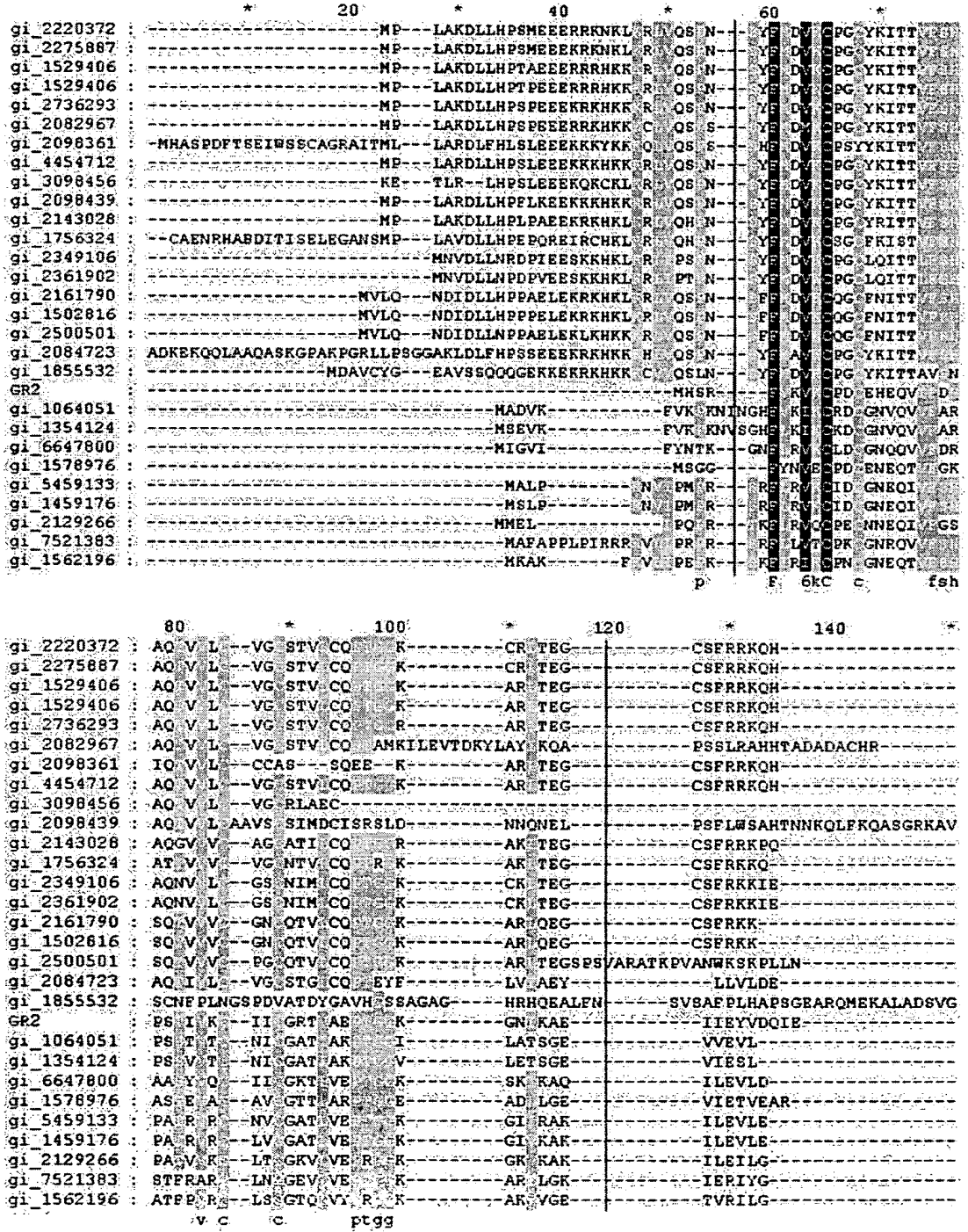


Figure 1. Multiple sequence alignment of *RS27_ARCFU* ending S27e with eukaryotic and archaebacterial homologs as detected by a PSI-BLAST search (Altschul et al. 1997). Note that eukaryotic sequences have an amino-terminal extension. The thin black lines at positions 56 and 120 denote the alignment used in the program ConSurf (Armon et al. 2001; Glaser et al. 2003), which maps sequence conservation to the surface of a representative structure (Fig. 5).

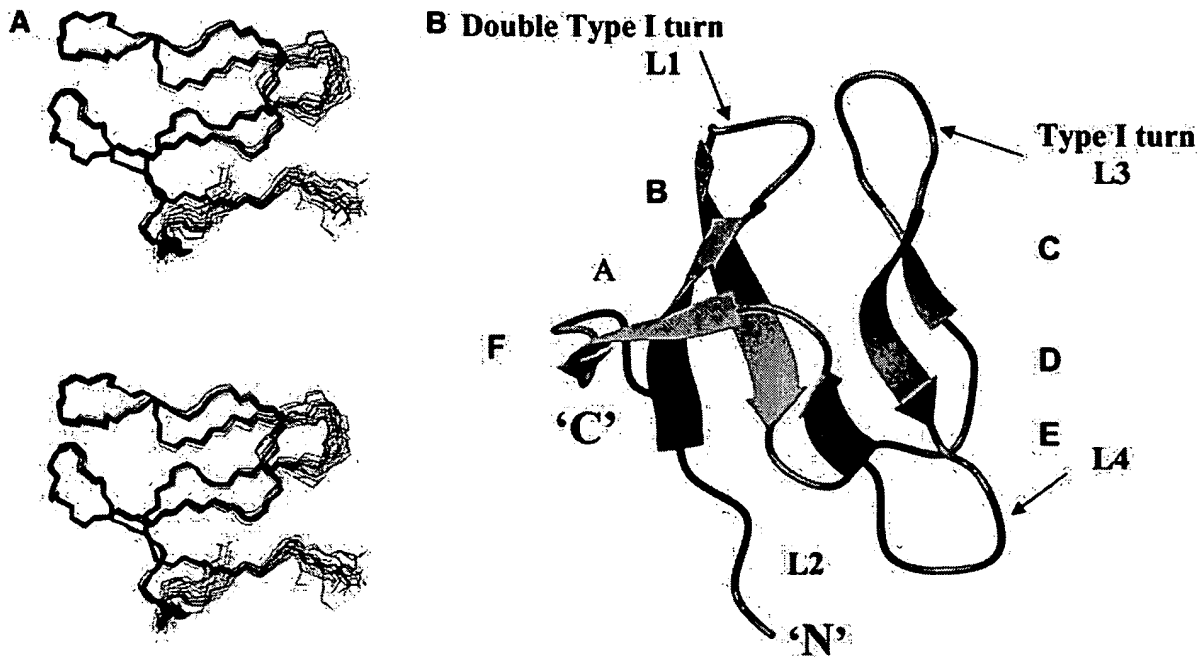


Figure 2. (*A*) The 20 DYANA conformers with the lowest residual DYANA target function chosen to represent the NMR solution structure of archaeal S27e are shown after superposition of the backbone heavy atoms N, C^{α} and C' for minimal RMSD. (*B*) Ribbon diagram of the backbone of the DYANA conformer with the lowest residual target function.

35–Val 36 and Ile 51–Glu 52, respectively. A double type I turn connecting β-strands A and B (loop L1) and a type I β-turn connecting strands C and D (loop L3) are structurally well defined (Figs. 3, 4) and located in close spatial proximity. Loops L1 and L3 contain the four cysteine residues of the zinc finger motif in a conformation required for zinc coordination, and the $^{13}C^{\beta}$ chemical shifts indicate that the cysteines are in reduced states (Atreya et al. 2000). The Φ and Ψ-dihedral angles of the double turn (L1; Fig. 4) are quite close to the ideal values (Hutchinson and Thornton 1994), whereas the Ψ dihedral angles of the type I turn (L3; Fig. 4) are somewhat more negative than expected (−80° and −51° instead of −30° and 0°). The two other loops, L2 connecting strands B and C, and L4 connecting strands D and E, are located on the opposite side of the β-sandwich with respect to L1 and L3, but on the same side as the amino- and carboxy-termini. L2 is moderately well defined, whereas the glycine-rich loop L4 (sequence: PTGGKG) is flexibly disordered in solution (Figs. 3, 4).

Novelty of S27e fold

First, no structural homologs were detected by searching S27e (Fig. 2) against the PDB using the programs CE (Shindyalov and Bourne 1998) or DALI (Holm and Sander 1993). Second, the structure of S27e has not yet been classified in CATH (Orengo et al. 1997), but a GRATH search yielded no structural homologs. Third, S27e could be assigned to the smaller protein class of the rubredoxin-like fold in the SCOP classification (Lo Conte et al. 2000). These proteins comprise a metal (zinc or iron)-bound fold, contain two CX(n)C motifs (in most cases $n = 2$) and the current SCOP classification encompasses 12 superfamilies. An automatic family pairwise search of the SCOP database gave by far the lowest e-value (3.5e-06) for the zinc beta-ribbon superfamily. However, interactive analysis reveals that the arrangement of secondary structure elements is unique in S27e. Taken together, these findings support our conclusion that S27e possesses a hitherto uncharacterized protein fold (Fig. 2).

Zinc finger motif

The threading program 3D-PSSM (Kelley et al. 1999) predicts that the sequence of S27e encoded in *RS27_ARCFU* is consistent with five structures deposited in the PDB (1FFK_W, 1GH9, 1PFT, 1QYP, and 1TFI), all of which have a potential zinc-binding site. Analysis of these five structures along with the structure of S27e using the structure alignment module in PrISM (Yang and Honig 1999) identifies 1GH9 as the closest structural neighbor of S27e, the RMSD is 3.2 Å over 50% of S27e (Fig. 5). 1GH9 corresponds to a 8.3-kD protein (gene *Mth1184*) from *Methanobacterium thermoautotrophicum* of unknown func-

Table 1. *Statistics for the final ensemble of 20 structures calculated for RS27_ARCFU*

Distance restraints	
Intraresidue	228
Sequential	170
Medium range ($1 < \|i - j\| \leq 5$)	79
Long range ($\|i - j\| > 5$)	192
Hydrogen bond restraints	0
Total	669
Dihedral restraints	
Φ	61
Ψ	30
Total	91
Number of structural constraints per residue	13
Number of long range constraints per residue	3
Residual target function (Å^2)	0.52 ± 0.04
Distance constraint violations per structure (>0.1 Å)	0
Dihedral constraint violations per structure (>5°)	0
RMSD relative to the mean coordinates (Å)	
All residues[a]	
Backbone heavy atoms	0.69 ± 0.17
All heavy atoms	1.10 ± 0.16
Regular secondary structure elements[b]	
Backbone heavy atoms	0.17 ± 0.05
All heavy atoms	0.54 ± 0.09
Ramachandran plot statistics (%)[c]	
Residues in most favored regions	76.3
Residues in additional allowed regions	20.2
Residues in generously allowed regions	2.6
Residues in disallowed regions	0.9

[a] RMSD values for residues 1–58.
[b] Residues 5–9, 16–20, 27–28, 35–38, 45–46, 49–52.
[c] Determined using the program PROCHECK-NMR (Laskowski et al. 1996).

tion, and consists of a β-sheet followed by an α-helix and an unstructured carboxyl terminus (Christendat et al. 2000). The β-sheet region contains a CXCX...XCXC sequence with Cys residues located in two proximal loops. However, zinc binding could not be experimentally verified, so that Christendat et al. (2000) hypothesize that there may be specificity for some other metal. Accordingly, although 1QXF and 1GH9 exhibit different folds, the putative metal-binding region of S27e is observed in 1GH9.

Identification of conserved structural motifs

A PSI-BLAST search (Altschul et al. 1997) identified 116 sequence homologs to *RS27_ARCFU*, the gene encoding S27e, and additional homologs are likely to be identified as the genome-sequencing projects continue. A multiple sequence alignment was constructed of 30 representative sequences, including 19 eukaryotic and 11 archaeal homologs (Fig. 1). Subsequently, the sequence conservation reflected by this alignment was mapped onto the S27e structure using the program Consurf (Armon et al. 2001; Glaser et al. 2003). Two main conserved features emerge for the S27 family of sequence homologs, the putative zinc-binding region (Fig. 5) and a surface patch of hydrophobic residues consisting of Phe 5, Ile 19, Phe 20, and Val 27 in S27e (Fig. 6). When the eukaryotic and archaeal sequences are analyzed separately, one finds that the eukaryotic sequences have clusters of conserved basic residues that are missing in the archaeal sequences. Finally, there is a highly conserved four-residue fragment, Pro 39–Thr 40–Gly 41–Gly 42 (positions 96–99 in Fig. 1) located in the disordered loop L4 on the same side of the β-sheets as the aforementioned hydrophobic cluster, but on the opposite side of the β-sheets with respect to the putative zinc-binding region (Fig. 2).

Homology modeling of human S27

The human S27 (hS27) protein has a 26-residue amino-terminal extension relative to archaeal S27e (Fig. 1). A helix is predicted (Rost 1996) for residues 10–20 of hS27 with high confidence, and residues 13–23 were identified as a nuclear localization sequence. Consistent with the latter finding, the assembly of eukaryotic ribosomes is initiated in the cellular nucleus (Fromont-Racine et al. 2003). A model for the full-length human S27 (Fig. 7) was thus constructed using a composite approach based on two structural templates as follows: (1) the amino-terminal segment of the ribosomal protein L37Ae (1FFK:W), which contains a nuclear localization sequence and a helix in the expected location, for the amino-terminal 26 residues, and (2) the S27e structure (Fig. 2) for the carboxy-terminal domain. Notably, the Verify3D profile (Bowie et al. 1991; Luthy et al. 1992) indicates that the entire hS27 was modeled with high quality, possibly beset by some inaccuracy as far as the relative orientation of amino-terminal helix and carboxy-terminal domain is concerned.

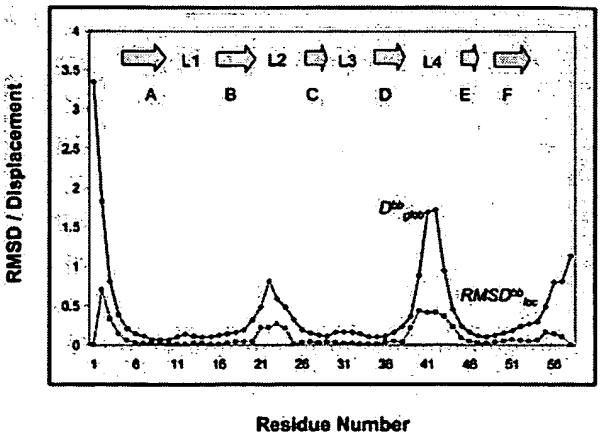


Figure 3. Global displacement (D^{bb}_{glob}) and local RMSD ($RMSD^{bb}_{loc}$) values calculated for the backbone heavy atoms N, C^{α} and C' from the ensemble of the 20 best DYANA conformers of S27e (Fig. 2) are plotted vs. the amino acid sequence.

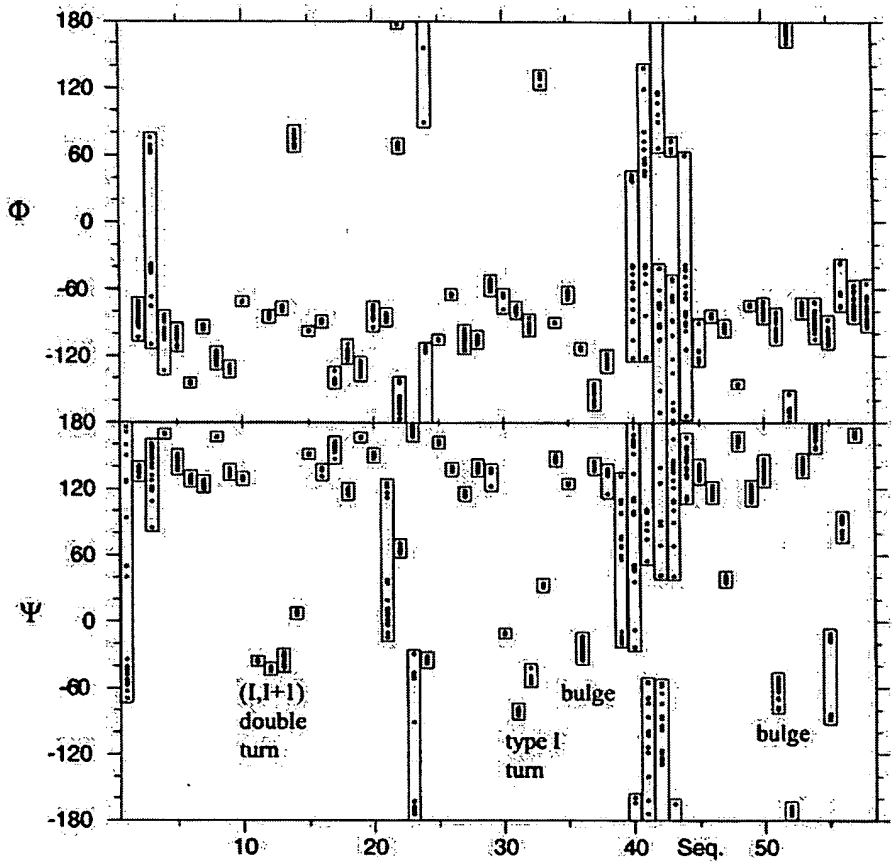


Figure 4. Plot of Φ and Ψ value ranges vs. the amino acid sequence for the 20 best DYANA conformers of S27e (Fig. 2; Table 1). The ranges for the residues forming the (I,I+1) double turn (L1), the type I turn (L3), and the bulges are indicated.

Intriguingly, the electrostatic surface potentials calculated for the archaeal S27e and the hS27 model are strikingly different (Fig. 8). S27e has two large patches of negative electrostatic potential (red), whereas the human S27 model is predicted to be overall positively charged (blue). Comparison of the electrostatic surface potentials of models of other archaeal and eukaryotic congeners (Fig. 1) shows that this is generally observed when comparing an archael and an eukaryotic S27 protein.

Conclusions

The conserved hydrophobic patch (Fig. 6) may well play a pivotal role for functioning of S27 proteins, be it in archaeal or eukaryotic cells. This finding suggests that the mode of operation of S27 proteins in the two kingdoms has common features. However, the dramatic change in the surface electrostatic potentials (Fig. 8) reveals the presence of ancient divergent evolution, which has led to either functional variations among archaeal and eukaryotic S27 proteins, or, assuming that the function remained invariant, to a concerted evolutionary change of the surface potential of proteins interacting with S27. The amino-terminal extension present in the eukaryotic S27 proteins appears to be required for nuclear ribosome assembly. In archaebacteria, the lack of compartmentation makes such a leader sequence evidently unnecessary. Most S27 proteins comprise a zinc finger (Fig. 1) capable of binding Zn or Fe ions. However, for YL37a (1FFK:W), which has been recruited to model hS27 (Fig. 7), three of the four cysteines could be replaced without noticeably affecting RNA binding or function in general (Dresios et al. 2002). Thus, it may well be that the zinc finger motifs that are highly conserved in the family of S27 proteins (Fig. 1) are indeed a vestigal structure characterizing ancient evolution of the ribosome before the divergence of archae and eubacteria lineages (Dresios et al. 2002).

Materials and methods

Protein purification

GR2 (*RS27_ARCFU*) was cloned, expressed, and purified following standard protocols to produce a uniformly (*U*) $^{13}C,^{15}N$-labeled protein sample. Briefly, the full-length GR2 gene from *Archae-*

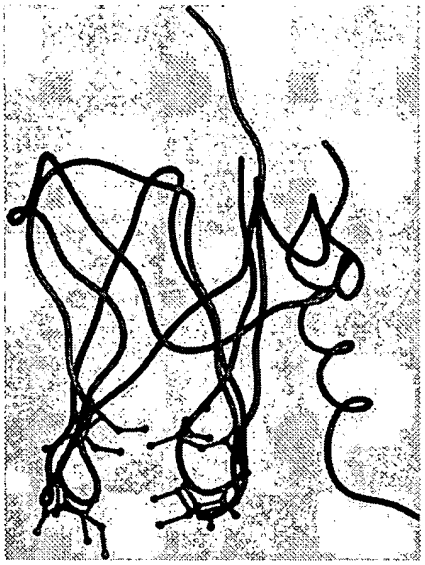


Figure 5. Structure superposition of S27e (red; Fig. 2) and 1GH9 (blue). The putative zinc-binding Cys residues are shown in ball-and-stick representation.

globus fulgidis was cloned into a pET21d (Novagen) derivative, yielding the plasmid pGR2-21. The resulting construct contains eight non-native residues at the carboxyl terminus (LEHHHHHH) that facilitate protein purification. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with pGR2-21, and cultured in MJ9 minimal medium containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as sole nitrogen and carbon sources (Jansson et al. 1996). *U*-^{13}C,^{15}N GR2 was purified using a two-step protocol consisting of Ni-NTA affinity (QIAGEN) and gel filtration (HiLoad 26/60 Superdex 75, Amersham Biosciences) chromatography. The final yield of purified *U*-^{13}C,^{15}N GR2 (>97% homogeneous by SDS-PAGE; 7.6 kD by MALDI-TOF mass spectrometry) was ~40 mg/L. In addition, a *U*-^{15}N and 5% biosynthetically directed fractionally ^{13}C-labeled sample was generated for stereospecific assignment of isopropyl methyl groups. The two samples, *U*-^{13}C,^{15}N and 5%^{13}C,*U*-^{15}N GR2, were prepared at concentrations of 1.0 mM in a 95% H_2O/5% D_2O solution containing 20 mM MES, 100 mM NaCl, 10 mM DTT, 5 mM $CaCl_2$, and 0.02% NaN_3 at pH 6.5.

NMR spectroscopy

All NMR data were collected at 25°C on a Varian INOVA 600 MHz spectrometer. Spectra were processed using the program PROSA (Guntert et al. 1992) or NMRPipe. (Delaglio et al. 1995) and subsequently analyzed using the program XEASY (Bartels et al. 1995). Resonance assignments were obtained from a minimal set of reduced-dimensionality NMR experiments (Szyperski et al. 2002) using 48 h of measurement time, including 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN (12 h), 3D HCCH COSY (20 h), and 3D HBCB(CGCD)HD (7 h), complemented by conventional 3D HNNCACB (5 h) and 3D HNNCO (3 h). Assignments were obtained for 97% of the backbone and $^{13}C^{\beta}$, and 98% side-chain chemical shifts. Stereospecific assignments of prochiral methyl groups of Val and Leu were obtained from 2D [^{13}C–^{1}H] HSQC for the 5% fractionally ^{13}C-labeled protein sample, which also supported the amino acid type identification (Neri et al. 1989; Szyperski et al. 1992). Chemical shifts have been deposited in the BMRB (accession no. 5682).

Structure calculation

NMR constraints were obtained from 3D HNNHA (Vuister and Bax 1993), from chemical shifts for residues in β-strands using the program TALOS (Cornilescu et al. 1999), ^{15}N- and ^{13}C-resolved

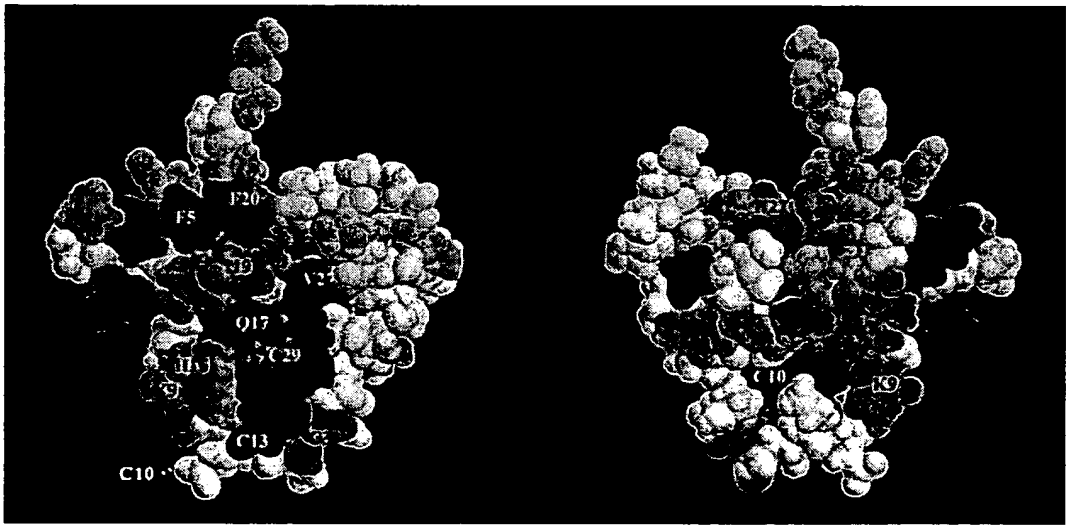


Figure 6. Multiple sequence alignment of Figure 1 analyzed by ConSurf (Armon et al. 2001; Glaser et al. 2003) using the NMR structure of archaeal S27e (Fig. 2). The most highly conserved residues are scored 9 and colored magenta, whereas the most variable sites are scored 1 and colored blue. Intermediately conserved residues are scored and colored on a graded scheme between these two extremes. The two views are rotated by 180 around the vertical axis with respect to each other. The hydrophobic patch formed by Phe 5, Ile 19, Phe 20, and Val 27 of S27e is readily apparent.

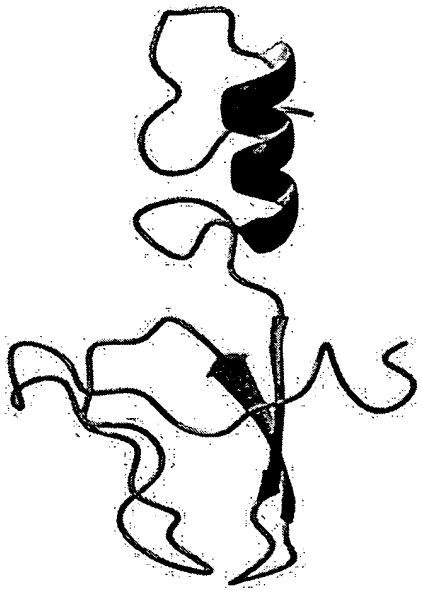


Figure 7. Ribbon representation of the homology model constructed for human S27 based on the structural templates of archaeal S27e (1QXF; Fig. 2) and the amino-terminal segment of the archaeal ribosomal protein L37Ae (1FFK:W).

[^{1}H-^{1}H] NOESY, and 2D [^{1}H,^{1}H]-NOESY (τ_m = 70 msec). NOE cross-peak assignments and volumes were obtained using the XEASY program. (Bartels et al. 1995). After obtaining the initial fold using the program DYANA (Guntert et al. 1997), the CANDID module within the CYANA program (Herrmann et al. 2002) was run in an iterative manner to obtain automated assignments and distance calibrations for the remaining NOEs. The program MolMol (Koradi et al. 1996) was used to analyze NMR structures and to generate figures.

Bioinformatics

The structure alignment module in PrISM (Yang and Honig 1999) was used to structurally superimpose S27e (Fig. 2) with the structures identified by searching the sequence of S27e against the PDB using the threading program 3D-PSSM (Kelley et al. 1999). PrISM identified 1GH9 as the closest structural neighbor with the following statistics:

structure 1	structure2	rmsd	score	%-ali	SeqID
1qxf (1–60)	1gh9:A(1–71)	3.2	385.3	0.500	0.267

An RMSD of 3.2 Å over 50% of 1QXF suggests a significant superposition.

The first iteration of the PSI-BLAST search (Altschul et al. 1997) for homologs of the gene *RS27_ARCFU* encoding S27e in the nonredundant (nr) protein sequence database used the BLOSUM62 substitution matrix (Henikoff and Henikoff 1992) and gap existence and extension penalties of 11 and 1, respectively. After the initial search, an e-value threshold of 0.001 was applied for including sequences to form the position-specific scoring matrix used in subsequent searches. The PSI-BLAST search converged after three iterations and yielded 116 sequence homologs.

The program ConSurf (Armon et al. 2001; Glaser et al. 2003) was used to map the sequence conservation reflected in the multiple sequence alignment of *RS27_ARCFU* (1QXF) with 29 representative sequence homologs, depicted in Figure 1, onto the *RS27_ARCFU* structure (Fig. 6). The level of sequence conservation at each position of the multiple sequence alignment is represented by coloring each residue in the structure according to the graded scheme described in the legend to Figure 6.

A high-quality homology model for residues 27–83 of human S27 was constructed using the S27e structure (Fig. 2) as the template. The amino-terminal segment of the ribosomal protein L37 (1FFK:W) was identified by 3D-PSSM (Kelley et al. 1999) as a good structural representation for the amino-terminal extension of human S27. A composite homology model for the entire human S27 sequence (Fig. 7) was constructed with the program Nest (Petrey et al. 2003) using the two templates, 1FFK:W and 1QXF, according to the following alignments:

N-terminus:

```
1FFK:    PTGR--FGPRYGLKIRVRVRDVEIKH
h_RS27: PLAKDLLHPSPEEEKRKHKKKRLVQS
```

C-terminus:

```
1QXF:    HSRFVKVKCPDCEHEQVIFDHPSTIVKCIIC
GRTVAEPTGGKGNIKAEIIEYVDQIE
h_RS27: PNSYFMDVKCPGCYKITTVFSHAQTVVLCVG
CSTVLCQPTGGKARL-EGCSFRRKQH
```

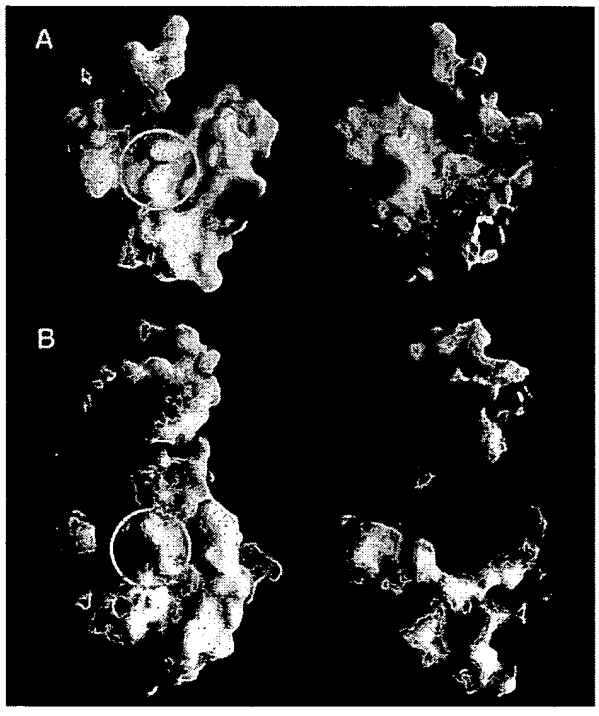


Figure 8. (*A*) Grasp (Nicholls et al. 1991) profiles representing the electrostatic surface potential of archaeal S27e. The scale of surface potentials is −5, 0, +5 kT/e with red (blue) corresponding to negative (positive) electrostatic potential. The *left* and *right* views are rotated by 180 about the vertical axis with respect to each other. (*B*) The same as in *A* for the homology model of human S27.

The homology model of hS27 (Fig. 7) scored well according to the structure evaluation program Verify3D (Bowie et al. 1991; Luthy et al. 1992). Electrostatic surfaces (Fig. 8) were calculated using the program GRASP (Nicholls et al. 1991).

Acknowledgments

This work was supported by the National Institutes of Health (P50 GM62413-01), the National Science Foundation (MCB 00075773 to T.S.), and the Center for Computational Research at UB. C.P. is indebted to the C.N.R.S. (France) for funding a sabbatical leave at the University at Buffalo.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation pf protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Armon, A., Graur, D., and Ben-Tal, N. 2001. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface-mapping of phylogenetic information. *J. Mol. Biol.* **307:** 447–463.

Atreya, H.S., Sahu, S.C., Chary, K.V.R., and Govil, G. 2000. A tracked approach for automated NMR assignments in proteins (TATAPRO). *J. Biomol. NMR* **17:** 125–136.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution. *Science* **289:** 905–920.

Bartels, C., Xia, T., Billeter, M., Guntert, P., and Wuthrich, K. 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6:** 1–10.

BaudinBaillieu, A., Tollervey, D., Cullin, C., and Lacroute, F. 1997. Functional analysis of Rrp7p, an essential yeast protein involved in pre-rRNA processing and ribosome assembly. *Mol. Cell. Biol.* **17:** 5023–5032.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into known three-dimensional structure. *Science* **253:** 164–170.

Brodersen, D.E., Clemons Jr., W.M., Carter, A.P., Wimberly, B.T., and Ramakrishnan, V. 2002. Crystal structure of the 30S ribosomal subunit from *Thermus thermophilus*: Structure of the proteins and their interactions with 16S RNA. *J. Mol. Biol.* **316:** 725–768.

Chan, Y.L., Suzuki, K., Olvera, J., and Wool, I.G. 1993. Zinc finger-like motifs in rat ribosomal-proteins S27 and S29. *Nucleic Acids Res.* **21:** 649–655.

Chothia, C., Levitt, M., and Richardson, D. 1977. Structure of proteins: Packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci.* **74:** 4130–4134.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., et al. 2000. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7:** 903–909.

Cornilescu, G., Delaglio, F., and Bax, A. 1999. Protein backbone angle restraints from a database for chemical shift and sequence homology. *J. Biomol. NMR* **13:** 289–302.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. 1995. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6:** 277–293.

Dresios, J., Chan, Y.-L., and Wool, I.G. 2002. The role of the zinc finger motif and of the residues at the amino terminus in the function of yeast ribosomal protein YL37a. *J. Mol. Biol.* **316:** 475–488.

Frankel, A.D. 2000. Fitting peptides into the RNA world. *Curr. Opin. Struct. Biol.* **10:** 332–340.

Fromont-Racine, M., Senger, B., Saveanu, C., and Fasiolo, F. 2003. Ribosome assembly in eukaryotes. *Gene* **313:** 17–42.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19:** 163–164.

Goldsmith-Fischman, S. and Honig, B. 2003. Structural genomics: Computational methods for structure analysis. *Protein Sci.* **12:** 1813–1821.

Güntert, P., Dotsch, V., Wider, G., and Wüthrich, K. 1992. Processing of multidimensional NMR data with the new software PROSA. *J. Biomol. NMR* **2:** 619–629.

Güntert, P., Mumenthaler, C., and Wüthrich, K. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273:** 283–298.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Herrmann, T., Güntert, P., and Wüthrich, K. 2002. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319:** 209–227.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.

Hutchinson, E.G. and Thornton, J.M. 1994. A revised set of potentials for β-turn formation in proteins. *Protein Sci.* **3:** 2207–2216.

Jansson, M., Li, Y.C., Jendeberg, L., Anderson, S., Montelione, G.T., and Nilsson, B. 1996. High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in *Escherichia coli*. *J. Biomol. NMR* **7:** 131–141.

Kelley, L.A., MacCallum, R.M., and Steinberg, M.J.E. 1999. Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. In *RECOMB 99, Proceedings of the third annual conference on computational molecular biology* (eds. S. Istrail et al.), pp. 218–225. The Association for Computing Machinery, New York.

Kim, S. and Szyperski, T. 2003. GFT NMR, a new approach to rapidly obtain precise high dimensional NMR spectral information. *J. Am. Chem. Soc.* **125:** 1385–1393.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.

Koradi, R., Billeter, M., and Wuthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14:** 51–55.

Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8:** 447–486.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **28:** 257–259.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* **356:** 83–85.

Montelione, G.T., Zheng, D., Huang, Y., Gunsalus, K.C., and Szyperski, T. 2000. *Nat. Struct. Biol.* **7:** 982–984.

Neri, D., Szyperski, T., Otting, O., Senn, H., and Wuthrich, K. 1989. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 Repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28:** 7510–7516.

Nicholls, A., Sharp, K.A., and Honig, B. 1991. GRASP-Graphical representation and analysis of surface properties. *Struc. Funct. Genet.* **11:** 281–296.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH - A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Petrey, D., Xiang, X., Tang, C.L., Xie, L., Gimpelev, M., Mitors, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., et al. 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* **53:** 430–435.

Ramakrishnan, V. and White, S.W. 1998 Ribosomal protein structures: Insights into the architecture, machinery and evolution of the ribosome. *Trends Biochem. Sci.* **23:** 208–212.

Revenkova, E., Masson, J., Koncz, C., Afsar, K., Jakovleva, L., and Paszkowski, J. 1999. Involvement of *Arabidopsis thaliana* ribosomal protein S27 in mRNA degradation triggered by genotoxic stress. *EMBO J.* **18:** 490–499.

Rost, B. 1996 PHD: Predicting one-dimensional protein structure profile based on neural networks. *Methods Enzymol.* **266:** 525–539.

Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janelle, D., Bhasan, A., Bartels, H., Agmon, I., Franceschi, F., et al. 2000. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102:** 615–623.

Shindyalov, I.N. and Bourne, P.E. 1998. Protein structural alignment by incremental combinatorial extention (CE) of the optimum path. *Protein Eng.* **11:** 739–747.

Szyperski, T., Neri, D., Leiting, B., Otting, G., and Wüthrich, K. 1992. Support

of ^{1}H NMR assignments in proteins by biosynthetically directed fractional ^{13}C-labeling. *J. Biomol. NMR* **2:** 323–334.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N.B., and Montelione, G.T. 2002. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. *Proc. Natl. Acad. Sci.* **99:** 8009–8014.

Takahashi, Y., Mitsuma, T., Hirayama, S., and Odani, S. 2002. Identification of the ribosomal proteins present in the vicinity of globin mRNA in the 40S initiation complex. *J. Biochem.* **132:** 705–711.

Vuister, G.W. and Bax, A. 1993. Quantitative *J* Correlation: A new approach for measuring homonuclear three-bond $J(H^{N}H^{\alpha})$ coupling constants in ^{15}N-enriched proteins. *J. Am. Chem. Soc.* **115:** 7772–7777.

Xia, Y.L., Arrowsmith, C.H., and Szyperski, T. 2002. Novel projected 4D triple resonance experiments for polypeptide backbone chemical shift assignment. *J. Biomol. NMR* **24:** 41–50.

Yang, A.S. and Honig, B. 1999. Sequence to structure alignment in comparative modeling using PrISM. *Proteins* **Suppl. 3:** 66–72.

Yee, A., Chang, X.Q., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G.M., Bhattacharyya, S., Gutierrez, P., et al. 2002. An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci.* **99:** 1825–1830.

Zheng, D., Huang, Y.J., Moseley, H.N.B., Xiao, R., Aramini, J., Swapna, G.V.T., and Montelione, G.T. 2003. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci.* **12:** 1232–1246.

EXHIBIT 25

High-quality homology models derived from NMR and X-ray structures of *E. coli* proteins YgdK and SufE suggest that all members of the YgdK/SufE protein family are enhancers of cysteine desulfurases

GAOHUA LIU,[1,2] ZHAOHUI LI,[1,3] YIWEN CHIANG,[1,4,5] THOMAS ACTON,[1,4,5] GAETANO T. MONTELIONE,[1,4,5] DIANA MURRAY,[1,3] AND THOMAS SZYPERSKI[1,2]

[1]The Northeast Structural Genomics Consortium
[2]Department of Chemistry, University at Buffalo, The State University of New York, Buffalo, New York 14260, USA
[3]Department of Microbiology and Immunology, Weill Medical College of Cornell University, New York, New York 10021, USA
[4]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854, USA
[5]Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey 08854, USA

(RECEIVED December 28, 2004; FINAL REVISION March 20, 2005; ACCEPTED March 24, 2005)

Abstract

The structural biology of proteins mediating iron-sulfur (Fe-S) cluster assembly is central for understanding several important biological processes. Here we present the NMR structure of the 16-kDa protein YgdK from *Escherichia coli*, which shares 35% sequence identity with the *E. coli* protein SufE. The SufE X-ray crystal structure was solved in parallel with the YdgK NMR structure in the Northeast Structural Genomics (NESG) consortium. Both proteins are (1) key components for Fe-S metabolism, (2) exhibit the same distinct fold, and (3) belong to a family of at least 70 prokaryotic and eukaryotic sequence homologs. Accurate homology models were calculated for the YgdK/SufE family based on YgdK NMR and SufE crystal structure. Both structural templates contributed equally, exemplifying synergy of NMR and X-ray crystallography. SufE acts as an enhancer of the cysteine desulfurase activity of SufS by SufE–SufS complex formation. A homology model of CsdA, a desulfurase encoded in the same operon as YgdK, was modeled using the X-ray structure of SufS as a template. Protein surface and electrostatic complementarities strongly suggest that YgdK and CsdA likewise form a functional two-component desulfurase complex. Moreover, structural features of YgdK and SufS, which can be linked to their interaction with desulfurases, are conserved in all homology models. It thus appears very likely that all members of the YgdK/SufE family act as enhancers of Suf-S-like desulfurases. The present study exemplifies that "refined" selection of two (or more) targets enables high-quality homology modeling of large protein families.

Keywords: YgdK; SufE; IscU; Fe-S cluster; NMR; homology modeling

Supplemental material: see http://www.proteinscience.org

Reprint requests to: Thomas Szyperski, Department of Chemistry, University of Buffalo, The State University of New York, 816 Natural Sciences Complex, Buffalo, NY 14260, USA; e-mail: szypersk@chem.buffalo.edu; fax: (716) 645-7338.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.041322705.

The *Escherichia coli* protein YgdK is a sequence homolog of the *E. coli* protein SufE, whose function as an enhancer of cysteine desulfurase activity in the primary *E. coli* Fe-S cluster assembly pathway has recently been demonstrated (Loiseau et al. 2003; Ollagnier-de-Choudens et al. 2003;

Outten et al. 2003). Cysteine desulfurases are pyridoxal 5′-phosphate (PLP)-dependent homodimeric enzymes that catalyze the conversion of L-cysteine to L-alanine and sulfane sulfur via the formation of a protein-bound cysteine persulfide intermediate on a conserved cysteine residue (Mihara and Esaki 2002). Cysteine desulfurases mobilize sulfur for biosynthesis, e.g., for Fe-S cluster assembly or thionucleoside biosynthesis (Mihara and Esaki 2002), and are found in almost all living organisms. However, the mechanisms for sulfur mobilization mediated by cysteine desulfurases are still unclear. The Gram-negative bacterium *E. coli* possesses three cysteine desulfurases, i.e., IscS, CsdA, and SufS (also known as CsdB), which also have significant sequence similarity with the *Azotobacter vinelandii* NifS desulfurase functioning in the process of nitrogen fixation. The genes encoding these enzymes are located at different loci and are coexpressed with different sets of accessory proteins.

Expression of bacterial cysteine desulfurases is generally regulated by operons that also control expression of several functionally related proteins. For example, the *suf* operon contains six genes encoding the proteins SufA, SufB, SufC, SufD, SufS, and SufE, while IscS is part of the similar Isc operon encoding "housekeeping" proteins required for Fe-S cluster biosynthesis. IscS interacts with IscU, a Zn-binding protein (Ramelot et al. 2004) also coded for by the Isc operon. IscU is proposed to function as a scaffold for the assembly of iron-sulfur clusters, whereby a sulfur atom is transferred from L-cysteine via IscS to IscU. Although sharing $<10\%$ sequence identity, the similarity of three-dimensional structure, surface features, and regulatory control suggests that IscU and SufE are homologous desulfurase enhancers, interacting with IscS and SufS, respectively (Goldsmith-Fischmann et al. 2004; Ramelot et al. 2004). CsdA and SufS share 45% sequence identity, but both exhibit 24% or less identity with NifS or IscS. Recently, it was shown that SufS and SufE, as well as IscS and IscU, form complexes, thus providing examples of two-component cysteine desulfurase enzymes (Smith et al. 2001; Urbina et al. 2001; Loiseau et al. 2003; Ollagnier-de-Choudens et al. 2003; Outten et al. 2003). The operon containing the gene of CsdA also encodes YgdK directly downstream of CsdA. Given the sequence homology between YgdK and SufE and between CsdA and SufS, it thus appeared quite likely that CsdA and YgdK form a complex similar to that formed by SufS and SufE.

YgdK and SufS share 35% sequence identity and were chosen for parallel structure determination by the Northeast Structural Genomics consortium (NESG) (http://www.nesg.org; target IDs ER75 for YgdK and ER30 for SufS). (Notably, IscU was also selected as a target protein by NESG (target ID IR24; Protein Data Bank [PDB] IDs 1Q48, 1R9P [Ramelot et al. 2004].) One goal of the Protein Structure Initiative (http://www.nigms.nih.gov/psi) is to experimentally solve at least one representative protein structure for each domain of several hundred domain sequence families. These structures serve as "structural templates" to homology-model the structures of other family members (Marti-Renom et al. 2000). The "leverage value" of a given structural template is estimated by assessing both the number of structures that can be modeled and the resulting quality of the models. Although results from the recent Critical Assessment of Protein Structure Prediction (CASP5) experiment suggest that sequence identity between target and template is not always a reliable indicator of the quality of a homology model (Tramontano and Morea 2003), it is generally acknowledged that the accuracy of a homology model scales with the sequence identity between modeled and template protein (Fiser et al. 2000). A recent study based on the Swiss-Model homology modeling server illustrates this point (Schwede et al. 2000). SwissModel was used to construct a set of 1200 "control" models, i.e., models for sequences with known structure based on templates with which they share between 25% and 95% identity. As expected, models based on alignments of higher sequence identity were structurally more similar to the actual structures than models based on alignments of lower sequence identity. For example, the percentage of models whose C^{α} atom coordinates were "within" an RMSD value of 2 Å to the experimentally determined structure was, respectively, 18% and 55% for sets with target-template sequence identities of 30%–39% and 50%–59% (Schwede et al. 2000).

Larger families of sequence homologs exhibit a larger range of sequence identity to the representative experimental template. For this reason, it is often necessary to select two (or more) experimental structures to obtain high-quality models for all members, especially when structural diversity is expected among the family members based on an examination of their sequences and secondary structure predictions. If the targets are selected judiciously, these multiple structures can provide a larger number of family members whose structures can be (more) accurately modeled. Iterative selection of multiple targets within a domain family, so as to provide proper coverage of the entire domain family, is a basic component of the target selection strategy of the NESG consortium (Liu and Rost 2002; Liu et al. 2004; Wunderlich et al. 2004). In the present study, we examine the coverage of sequence space by the structures of YgdK (147 residues) and SufE (138 residues), which belong to NESG consortium target cluster 8976 (http://www.nesg.org).

Here we report (1) the high-quality NMR solution structure of YgdK (PDB ID 1NI7), (2) its comparison

with the 2.0 Å X-ray crystal structure of SufE (PDB ID 1MZG) that was solved in parallel by Goldsmith-Fischmann et al. (2004), and (3) a thorough search for other structurally similar proteins in the PDB. High-quality homology models were then calculated for 68 out of a family of 70 sequence homologs comprising YgdK and SufE (the "YdgK/SufE" family), and a "leverage analysis" is presented. In conjunction with a homology model for CsdA, which was derived from the crystal structure of SufS, the conservation of structural motifs in the set of homology models allowed us also to identify key features for the putative YgdK–CsdA complex formation. The modeling yields novel insights into the structural biology of two-component desulfurases involved in Fe-S cluster assembly.

Results and Discussion

Resonance assignments

The approximate isotropic overall rotational correlation time for protein YdgK was inferred from ^{15}N $T_1/T_{1\rho}$ nuclear spin relaxation time ratios (Szyperski et al. 2002). In agreement with a molecular mass of 16 kDa, a value of ~8.5 nsec was obtained, which shows that YgdK is monomeric in solution. This enabled collection of high-quality NMR spectra.

Following the protocol described previously (Szyperski et al. 2002), reduced-dimensionality (RD) $^{13}C/^{15}N/^{1}H$ triple resonance NMR spectroscopy, complemented by heteronuclear resolved [^{1}H,^{1}H]-NOESY (Cavanagh et al. 1996), was used for the resonance assignment of $^{13}C/^{15}N$- and ^{15}N-labeled YgdK. Complete assignments were obtained for backbone and $^{13}C^{\alpha}$ chemical shifts (excluding the "His tag") with the sole exception of (1) the $^{13}C'$ resonances of the residues that precede Pro, and Met 1, Thr 2, and Gly 85; and (2) the backbone amide resonances of Met 1, Thr 2, Asn 3, and Arg 86. Notably, the detection of strong sequential $d_{\alpha\delta}$ or $d_{N\delta}$ NOEs showed that all prolyl residues (4, 10, 26, 45, 112) adopt a *trans*-conformation (Wüthrich 1986). Complete assignments were also obtained for (1) the aliphatic side-chain resonances, except those of Met 1 and Thr 2; (2) all aromatic side-chain resonances (except for H^{ζ} of Phe 6, 11, 79, and $H^{\epsilon 1}$ of His 78); and (3) all side-chain amide groups of Asn and Gln residues. Furthermore, the HN^{ε} of five Arg residues (21, 35, 64, 86, 89) as well as the hydroxyl protons of three Thr residues (13, 97, 144) and two Ser residues (83, 137) could be assigned. Overall, 98% and 97% of the, respectively, routinely assigned backbone and side-chain shifts (see Table 1 footnote) were obtained and deposited in the BioMagResBank (accession code 5630).

Table 1. *Statistics of YgdK NMR structure determination*

Completeness of resonance assignments (%)	
Backbone[a]	98%
Side chains[b]	97%
Stereospecific assignments[c]	
$^{\alpha}$CH2 of glycines	83%
$^{\beta}$CH2	70%
Val and Leu isopropyl groups	80%
Conformationally restricting distance constraints	
Intraresidue	571
Sequential	701
Medium range	665
Long range	623
Total	2560
Dihedral angle constraints	
ϕ	213
ψ	102
Number of constraints per residue	19.6
Number of long-range constraints per residue	4.2
DYANA target function (Å^2)	0.67 ± 0.09
Average RMSD to the mean DYANA coordinates (Å)	
Regular secondary structure elements, backbone heavy atoms	0.53 ± 0.08
Regular secondary structure, all heavy atoms	0.88 ± 0.08
Residues 7–147, backbone heavy atoms N, C$^{\alpha}$, C'	0.72 ± 0.12
Residues 7–147, all heavy atoms	1.13 ± 0.12
Heavy atoms of molecular core[d]	0.85 ± 0.08
Ramachandran plot summary for residues 7–147 (%)	
Most favorable regions	83
Additionally allowed regions	16
Generously allowed regions	1
Disallowed regions	0
Average number of distance constraint violation per DYANA conformer (Å)	
0.2–0.5	0.7
>0.5	0
Average number of dihedral angle constraint violations per DYANA conformer (degrees)	
0–10	0.1
>10	0

[a] The N-terminal NH_3^+, Pro N, and carbonyl C' before Pro residues were not considered to calculate the fraction.
[b] Lys NH_3^+, Arg NH_2, side-chain carbonyl, and aromatic quaternary carbons were not considered to calculate the fraction.
[c] Relative to pairs with nondegenerate chemical shifts.
[d] Includes 47 residues: residues 10, 11, 15, 16, 17, 20, 24, 27, 30, 37, 40, 44, 47, 55, 58, 65, 66, 67, 79, 80, 87, 88, 91, 92, 93, 94, 95, 96, 97, 98, 99, 104, 105, 107, 112, 113, 115, 116, 119, 121, 125, 133, 136, 140, 141, 144, 147.

Structure determination of YgdK

A total of 2560 conformationally constraining NOE distance constraints were derived from 3D ^{15}N- and ^{13}C-resolved [^{1}H,^{1}H]-NOESY. In addition, $^{3}J_{HN\alpha}$ scalar couplings measured in 3D HNNHA (Vuister and Bax 1993) yielded 110 ϕ-angle constraints, and 206 backbone dihedral angle constraints were derived from chemical shifts as described using the program TALOS (Cornilescu et al. 1999). Using the FOUND and GLOMSA modules

of the program DYANA (Güntert et al. 1997), this set of experimental constraints provided stereospecific assignments (Table 1) for 10 Gly α-methylene proton pairs (83% of the pairs with nondegenerate chemical shifts), 52 β-methylene proton pairs (70% of the pairs with nondegenerated shifts), 19 more peripheral methylene proton pairs, and 20 isopropyl groups (the seven Val residues and 13 of the 21 Leu residues, i.e., 80% of the isopropyl methyl groups with nondegenerate chemical shifts).

An illustration of the quality of the YgdK structure is afforded by Figure 1a, which shows the polypeptide backbone of the 20 DYANA conformers selected to represent the solution structure after superposition of the backbone heavy atoms of the regular secondary structure elements. The small size and number of residual constraint violations show that the constraints are well satisfied in the set of 20 conformers (Table 1), and average RMSD values relative to the mean coordinates of 20 DYANA conformers of 0.72 Å for the backbone and of 1.13 Å for all heavy atoms are indicative of a high-quality NMR solution structure. Moreover, plots of local backbone RMSD values and global backbone displacements versus the sequence (Fig. 2a) show that all regular secondary structure elements are very well defined. Increased local disorder is observed only for the N-terminal hexapeptide segment, and the loop regions comprising residues 59–62 and 122–125. Comparison of RMSD values and displacements shows that these loops exhibit both local and global disorder. The coordinates of the YgdK NMR structure have been deposited in the P DB (ID 1NI7).

The high quality of the YgdK structure is further evidenced by (1) the large fraction of stereospecific assignments that have been obtained for the β-methylene and the Val and Leu isopropyl moieties (Table 1); (2) the fact that 83% of all ϕ and ψ dihedral angles are located in the "most favorable regions" of the Ramachandran plot (Table 1), while none of the residues is located in the "disallowed regions"; (3) an average G factor of −0.41 calculated for the backbone using the program Procheck (Laskowski et al. 1993, 1996); and (4) the identification of a large set of (subtle) helical capping motifs (see below). Highest-quality NMR solution structures have previously been assumed (Billeter 1992; Clore and Gronenborn 1998) to be comparable to 2.0–2.5 Å X-ray crystal structures. It thus appears that the quality of the NMR structure of YgdK and the 2.0 Å X-ray crystal structure of SufE (PDB ID 1MZG) exhibit comparable accuracy. Evidently, having two accurately determined structures is a favorable starting point for the desired high-quality homology modeling of a larger family of sequence homologs. In view of the homology modeling, it is particularly important that the molecular core of YgdK is very well defined by the NMR data, as reflected by an RMSD value of 0.84 Å for all heavy atoms of the core (Table 1).

Fold of YgdK

YgdK exhibits an α + β tertiary fold (Fig. 1b) that is composed of six α-helices, I to VI, which comprise residues 17–25, 30–44, 86–99, 104–110, 112–120, and 127–147, and a three-stranded anti-parallel β-sheet with strands A to C comprising, respectively, residues 56–58, 65–68, and 80–83 (Fig. 1b). In addition, a short helix II′ (residues 48–55) is present immediately N-terminal to strand A. Helices III and VI form a "coiled-coil" motif, and both helices are attached to

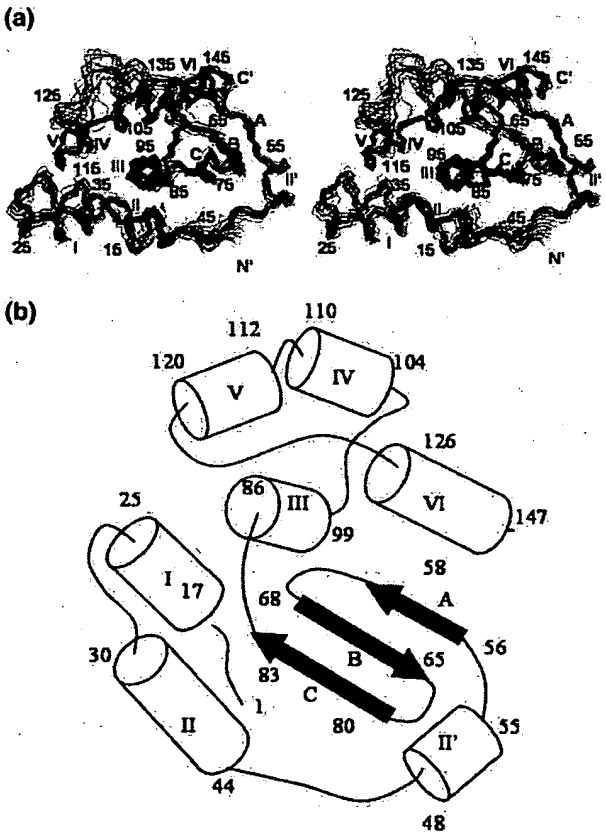


Figure 1. (*a*) Stereo view of the backbone of the 20 DYANA conformers representing the NMR solution structure of YgdK structure, after superposition of the backbone heavy atoms N, C^{α}, and C′ of the regular secondary structure elements for minimal RMSD. The polypeptide chain termini, the α-helices (I–VI, and II′), and the β-strands (A, B, and C) are indicated (see also Fig. 3). α-Helix I, residues 17–25; α-helix II, 30–44; α-helix II′, 48–55; α-helix III, 86–99; α-helix IV, 104–110; α-helix V, 112–120; α-helix VI, 127–147; β-strand A, 56–58; β-strand B, 65–68; β-strand C, 80–83. (*b*) Arrangement of regular secondary-structure elements identified for YgdK. The start and the end of the regular secondary-structure elements and the polypeptide chain ends are marked with their sequence location.

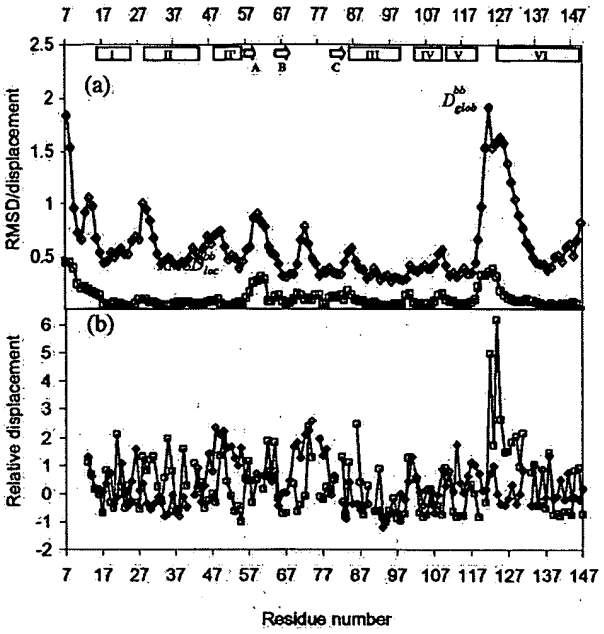


Figure 2. (*a*) Plots vs. the amino acid sequence of the mean global backbone displacements per residue, D_{glob}^{bb} (diamond), and the mean local RMSD, $RMSD_{loc}^{bb}$ (square), of the 20 DYANA conformers relative to the mean NMR structure calculated after superposition of the backbone heavy atoms N, C^{α}, and C′ of the regular secondary structures for the minimal RMSD. The local RMSD values are calculated for the tripeptide segments and plotted at the position of the central residue. (*b*) Relative displacements for the NMR structure of YgdK structure (square) and X-ray crystal structure SufE (diamond) calculated as described in Billeter (1992). For the NMR structure, the relative displacement, D_r(NMR), was calculated according to $D_r(\text{NMR}) = (D - \langle D \rangle)/\Delta D$, where $\langle D \rangle$ and ΔD are, respectively, the average displacement and standard deviation of displacement of each residue after superposition of the backbone heavy atoms N, C^{α}, and C′ for minimal RMSD. For the X-ray crystal structure, the relative displacement, D_r(X-ray), was calculated according to $D_r(\text{X-ray}) = (\sqrt{B} - \langle \sqrt{B} \rangle)/\Delta\sqrt{B}$, where $\langle \sqrt{B} \rangle$ and $\Delta\sqrt{B}$ are, respectively, the average and standard deviation of the square root of crystallographic *B*-factors of the backbone heavy atoms of a given residue.

one side of the β-sheet with helix III being oriented parallel to strand C. Helix IV is oriented approximately anti-parallel to helix III. The remaining helices I, II, and V surround helix III. As a result, helix III is largely buried in the protein's core (Fig. 3). The CATH protocol (Orengo et al. 1997; Pearl et al. 2000) assigns YgdK to the "α-β" "fold" class having a "two-layer sandwich" architecture, and YgdK is, for obvious reasons, assigned as a "SufE-like" fold in the class of α and β proteins in the SCOP classification.

Helix capping in YgdK

The high quality of the YgdK NMR solution structure allows one to identify capping motifs of helices, which play an important role for stabilizing the helices themselves as well as supersecondary structures (Aurora and Rose 1998). When searching for N-terminal capping interactions (Harper and Rose 1993), the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts afford a tentative identification of caps (Gronenborn and Clore 1994): The $^{13}C^{\alpha}$ chemical shift of the N-capped residue exhibits a 1–2-ppm upfield shift, while a downfield shift of 1–4 ppm is registered for its $^{13}C^{\beta}$ shift. N-capping interactions were inferred for helices I to IV from the chemical shifts. Inspection of the three-dimensional structure provides insight at atomic resolution. In helix I, Thr 16 is the N-cap residue and forms a capping box with Thr 19 as residue N3 (following the nomenclature of Aurora and Rose 1998): The hydrogen bonds Thr 16 HN–Thr 19 OG1 and Thr 16 OG1–Thr 19 HN are formed. The N terminus of helix I is further stabilized by N′–N4 hydrophobic interaction (a "hydrophobic staple motif") involving Val 15 and Leu 20. Its C terminus is capped by hydrophobic contacts involving the side chains of Phe 24 and Leu 27 as well as those of Thr 23 and Leu 27. Helix II is likewise stabilized by an N-terminal capping box: HN of the N-cap residue Gln 29 is hydrogen-bonded with the side-chain carboxylate of N3 residue Asn 32, and NH of Asn 32 forms a hydrogen bond with the side-chain carboxyl oxygen of Gln 29. Helix II′ exhibits a N′–N3 N-terminal hydrophobic capping motif coined the "h-xpxhx" motif (Aurora and Rose 1998), in which the methyl groups of Leu 47 are in close contact with those of Leu 51. Helix IV is N-terminally stabilized by a hydrogen bond formed between the amide proton of the N3 residue Glu 106 and the side-chain hydroxyl group of the N-cap residue Thr 103, and additional stabilization is due to the interaction of the side chains of Lys 102 and Glu 106. The C terminus of helix V exhibits a "Schellman motif," which involves both a hydrogen bond formed between Leu 121 HN and Phe 116 O′ and hydrophobic interactions between the side chains of these two residues.

Molecular core of YgdK

The tertiary fold of YgdK is stabilized by the formation of a molecular core involving side chains from all regular secondary structure elements except helix II′ (Fig. 3b). Notably, helix III is nearly entirely embedded in the core. As a result, most residues of helix III are hydrophobic, and there are only two charged residues, i.e., Arg 86 and Arg 89, among the 14 residues forming helix III. In fact, except for the two Arg residues, all side chains of helix III (Ile 87, Val 88, Leu 91, Leu 92, Ala 93, Val 94, Leu 95, Leu 96, Thr 97, Ala 98, and Val 99) are located in the interior of the protein and participate in hydrophobic contacts in the core. This is

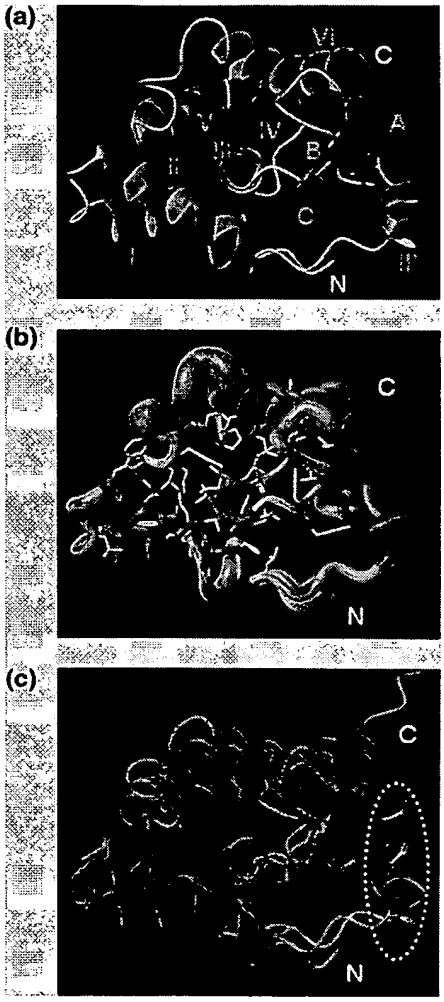


Figure 3. (*a*) Ribbon drawing of the DYANA conformer of YgdK (shown in the standard orientation of Fig. 1a) that exhibits the smallest RMSD value relative to the mean coordinates after superposition of the backbone heavy atoms of the regular secondary structure elements (Fig. 1b). The seven helices, I to VI and II′, are shown in red and yellow; the β-strands A, B, and C are depicted in cyan; and other polypeptides are displayed in gray. (*b*) Backbone of protein YgdK and the side chains forming the molecular core in the standard orientation of Figure 1a. For the presentation of the backbone, a spline function was drawn through the C^{α} positions; the thickness of the cylindrical rod is proportional to the mean of the global displacements of the 20 DYANA conformers calculated after superposition as described in Figure 1. The helices are shown in red, the β-stands are depicted in cyan, other polypeptide segments are displayed in gray, and the side chains of the molecular core are shown in yellow. (*c*) Ribbon drawings (in the standard orientation shown in Fig. 1a) of the DYANA conformer with the lowest target function value (Table 1) of YgdK (cyan) and the X-ray crystal structure SufE (magenta) after superposition of the C^{α} coordinates for minimal RMSD. The polypeptide segment of helix II′ in YgdK, which is corresponding to a coil region in SufE, is encircled (see text).

also reflected by solvent-exposed surfaces being below 10% for all theses residues. The side chains of helix III serve as a "nucleus" for formation of the molecular core interacting with the side chains of (1) Pro 10, Phe 11, and Val 15 located in the turn preceding helix I; (2) Ala 17, Leu 20, and Phe 24 of helix I; (3) Leu 27 located in the loop connecting helices I and II; (4) Trp 30, Leu 37, Leu 40, and Leu 44 of helix II; (5) Leu 58 of strand A; (6) Val 65 and Leu 67 of strand B; (7) Phe 79 and Phe 80 of strand C; (8) Ala 104, Ala 105, and Leu 107 of helix IV; (9) Pro 112, Leu 113, Leu 115, Phe 116, and Leu 119 of helix V; (10) Leu 121 and Leu 125 located in the loop connecting helices V and VI; and (11) Leu 133, Leu 136, Ile 140, Ile 141, Thr 144, and Val 147 of helix VI. Helix II′ is positioned by hydrophobic contacts with strand C. Those involve Ala 55 of helix II′, Leu 47 located in the loop connecting helix II and helix II′, and Trp 66 of strand B and Phe 80 of strand C. As a result of this tight network of mostly hydrophobic interactions, the molecular core of YgdK is very well defined in the NMR structure: The average RMSD value of all heavy atoms of the molecular core relative to the mean coordinates is 0.85 Å (Table 1), which is only slightly larger than the corresponding value obtained for the backbone heavy atoms alone.

Comparison of YgdK NMR and SufE crystal structure

The structure of SufE was solved in parallel by X-ray crystallography by Goldsmith-Fischmann et al. (2004). YgdK (147 residues, PDB ID 1NI7) and SufE (138 residues, 1MZG) exhibit 35% amino acid sequence identity, which clearly suggests that the two proteins adopt the same fold. Indeed, the RMSD calculated between the mean C^{α} coordinates of YgdK and the C^{α} coordinates of SufE is 2.5 Å. Figure 3c affords a visual impression of the global structural similarity. Moreover, inspection of the backbone dihedral angles ϕ and ψ shows that the two structures are also locally rather similar, with the exception of the polypeptide segment of helix II′ in YgdK: The corresponding segment in SufE does not adopt a helical conformation. As a result, a local RMSD value of 4.3 Å is calculated between the mean C^{α} coordinates of residues 45–54 of YgdK and the C^{α} coordinates of residues 35–44 of SufE after global superposition of all C^{α} coordinates of the two structures. This structural variation might be related to details of functional differences between the two proteins. An additional structural variation is manifested in somewhat different orientations of helices I and VI. Compared to YgdK, helices I and V are slightly shifted away from helices III and VI in SufE (Fig. 3c).

Internal motional modes are assumed to play an important role for protein function (for a recent review,

see Palmer 2001). We have thus calculated (Fig. 2b) the relative displacements, as described by Billeter (1992), for the backbone heavy atoms in YgdK and SufE structures. For the NMR structure, the relative displacement is derived from local RMSD values, while B-factors are recruited for the crystal structure. Relative displacements turn out to be rather similar throughout the polypeptide backbone for both proteins. However, the flexible disorder found for segment 122–125 in YgdK is apparently not observed for the corresponding segment 113–115 in SufE (Fig. 2b).

Structural similarity search for identifying potential homologs of YgdK

The program CE (Shindyalov and Bourne 1998) was used to search the PDB (Berman et al. 2000) for proteins other than SufE and IscU that are structurally similar to YgdK (see Supplemental Material for a detailed summary of the findings, including Table S1). In all cases, we found that (1) *z*-scores are at the threshold of being significant (i.e., around 4), (2) RMSD values calculated between the YgdK structure and these potential homologs are fairly high (between 3.5 Å and 4.0 Å), and (3) only ~60% of the YgdK structure can be aligned with any of the potential homologs (Table S1). Use of the programs Dali (Holm and Sander 1995) and PrISM (Yang and Honig 1999) supports the view that the structural similarity of currently known protein structures with YgdK is low and likely not significant (see Supplementall Material).

Homology modeling with YgdK and SufE: A "leverage analysis"

PSI-BLAST (Altschul et al. 1997) searches with the sequences of either YgdK or SufE against the nonredundant protein sequence database yielded the same set of 70 sequence homologs (including YgdK and SufE). Of these proteins, 66 are coded by prokaryotic genomes, while the remaining four are SufE-like domains of longer eukaryotic sequences. As a key result of this study, we were able to construct high-quality homology models for all 68 of these putative homologs of YgdK/SufE (see Materials and Methods). In order to assess the quality of the modeling protocol, we built a model of YgdK based on its alignment to SufE and using the structure of SufE as a template and vice versa. The RMSD value calculated between the C^{α} coordinates of the two models and their corresponding experimental structures is 2.6 Å in both cases, i.e., nearly exactly as large as the RMSD values calculated between the two experimental structures. We thus conclude that our models are accurate "within" 2.5–3 Å.

In Figure 4, the 68 homology-modeled YgdK/SufE family members are grouped according to their difference in sequence identity relative to YgdK and SufE (chart on the left of Fig. 4), as well as according to their sequence identity relative to both YgdK and SufE (chart on the right of Fig. 4). For more than one-half of the modeled proteins, one of the two template sequences exhibits 5% or more sequence identity to a given homolog than the other one, and 54 homologs are >30% identical to either YgdK or SufE, while 14 homologs are <30% identical to either of the two template sequences. Solving the structures of both YgdK and SufE enabled us to structurally characterize with high accuracy 68 naturally occurring proteins by using computational methods. Models for the 54 sequences with higher than 30% identity to either YgdK or SufE were constructed

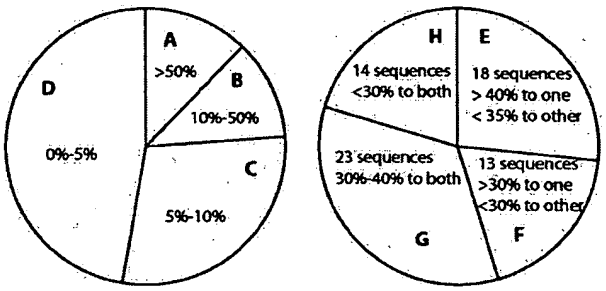


Figure 4. Summary of sequence identities between YgdK/SufE sequence homologs and YgdK and SufE, revealing the "leverage" of homology models obtained when using the experimentally determined structures of YgdK and SufE as templates. The 68 YgdK/SufE sequence homologs are grouped according to their difference in sequence identity to one structure over the other (*left* panel; A–D) and according to their sequence identity to both of the two structures (*right* panel; E–H). A, Eight sequence homologs are 50% more identical to one structure than the other; B, eight sequence homologs are between 10% and 50% more identical to one structure than the other; C, 20 sequence homologs are between 5% and 10% more identical to one structure than the other; D, 32 sequence homologs are between 0% and 5% more identical to one structure than the other; E, 18 sequences have >40% identity (40% to 99%) to one structure and <35% to the other; F, 13 sequences have >30% identity to one structure and <30% to the other; G, 23 sequences have between 30% and 40% identity to both structures; H, 14 sequences have <30% identity to both structures. (SWISS-PROT/TrEMBL primary AC: Q46926) and SufE (SWISS-PROT/TrEMBL primary AC: P76194). SWISS-PROT/TrEMBL primary accession numbers of the YgdK/SufE homologs are 26247929, 15802091, 24113068, 16760533, 16764724, 16122621, 22125831, Q9EXP1, 24324015, 24114095, 26249217, 16761763, 16121327, Q9KPQ6, 27365156, 23040025, P74523, 22298633, 17231005, 23124053, 23122888, Q9CME7, P44156, 23470497, 23108626, 26988260, 23467330, 23133309, 23060240, 22963827, 24375284, 21290341, 23028414, 15888257, Q985J5, 23501463, 27378180, Q9A9C7, 17987642, Q52967, 642658, 23105105, Q52742, 22958630, O96155, Q9PEN4, 24215721, 22994520, 22997695, 23135881, 21231690, 21243089, Q9HXX2, 23004440, 23131684, O65584, Q9Z9B0, Q9JSJ9, Q9K208, O84327, Q9PK71, 22974880, Q9RXU0, P96889, 23480445, 23593281, Q9FXE3, and Q9FGS4.

using the structure to which it had higher sequence identity as the template. Either structure would have served as a template in these cases, though, as expected, significantly better models were obtained when the higher similarity template was used. Importantly, it was also possible to build high-quality models for the 14 sequences that had <30% identity to both YgdK and SufE. In these cases, significantly better models could be constructed using one structure over the other as a template with a query/template alignment extracted from the multiple sequence alignment of all homologs. In addition, the alignments for modeling the sequences in this group were manually edited based on information from secondary structure predictions. Hence, the availability of both structures allowed us to model well this set of low-similarity sequences.

Three sets of 68 homology models for the YgdK/SurfE family were used for the leverage analysis: (1) a set generated with YgdK as template, (2) a set generated with SurfE as template, and (3) a set generated by using the higher similarity structure as template. All models score well according to the structure evaluation programs Verify3D (Bowie et al. 1991; Luthy et al. 1992) and Prosa2 (Sippl 1993), i.e., the models yielded Verify3D profiles that are usually obtained for high-quality experimentally determined structures. However, the models of the third set scored higher than those from the two single template sets. Notably, among the third set of 68 homology models, 34 models each were based on both the SufE and Ygdk structures. Hence, both experimental structures contributed equally to the excellent leverage in terms of both the number of modeled structures and their predicted quality. This finding (1) supports the view that the accuracy of YgdK NMR and SufE crystal structures is comparable, and (2) evidences a case of strong synergy of NMR and X-ray structure analysis for making available high-quality homology models of larger families of sequence homologs in structural genomics. The homology models are available online (http://maat.med. cornell.edu/nesg/er75_model.html).

Conserved surface patches identified in the YgdK/SufE protein family

Figure 5 depicts the sequence conservation among the YgdK/SufE homologs as well as the corresponding

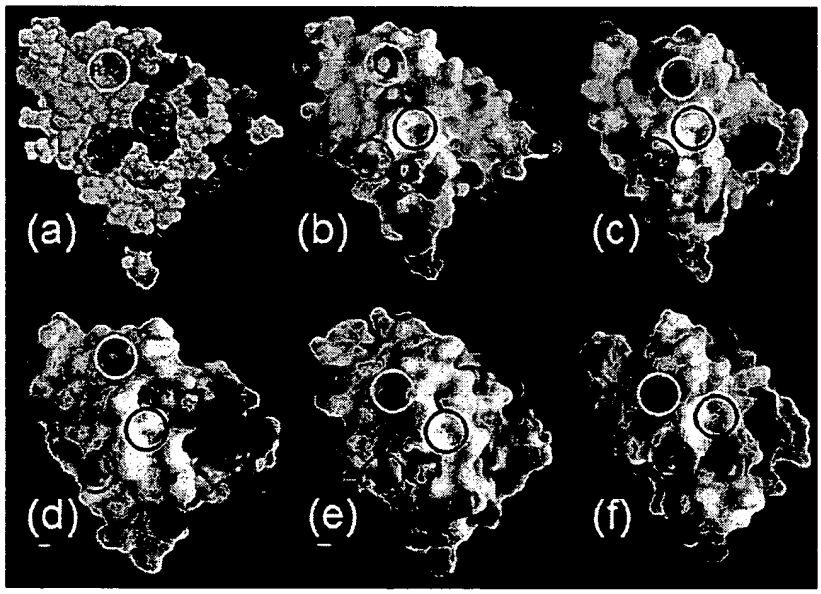


Figure 5. Conservation of surface features among sequence disparate YgdK/SufE homologs. All protein structures/models are shown in the same orientation. (*a*) Conserved surface residues identified using the program ConSurf (Glaser et al. 2003) with the YgdK NMR structure and the 69 sequence homologs of YgdK (see text). Two conserved charged residues and the cysteine implicated in sulfur transfer are labeled and circled to facilitate comparison with the surface electrostatic potential distribution calculated with the program GRASP (Nicholls et al. 1991), as displayed for YgdK in *b*, and homology models of four sequence homologs: Q9FXE3 (*c*), P96889 (*d*), Q9K208 (*e*), and P74523 (*f*). The three residues labeled in *a* are present in all models except Q9K2O8, for which a Lysine is at the position of Glu 84 (green circle, *e*). The percent pairwise sequence identities based on sequence (structure) alignments are: Q9FXE3/YgdK 25(21); P96889/YgdK 29(24); Q9K208/YgdK 17(15); P74523/YgdK 29(24); P96889/Q9FXE3 23(21); Q9K208/Q9FXE3 18(14); P74523/Q9FXE3 29(29); Q9K208/P96889 15(17); P74523/O96889 24(21); P74523/Q9K208 23(22). Both the multiple sequence alignment and the structure-based multiple sequence alignment were constructed in PrISM (Yang and Honig 1999).

conservation of surface features among the structure of YgdK and four models of sequence-disparate homologs, i.e., none of the sequences whose models are presented in Figure 5 shares higher than 30% identity with any of the other four proteins shown (see figure legend). Nonetheless, there are highly conserved residues (depicted in maroon in Fig. 5a) that support conserved surface properties across the family, most notably Cys 61, which is implicated in sulfur transfer (Loiseau et al. 2003; Ollagnier-de-Choudens et al. 2003; Outten et al. 2003), and a basic (Arg 129) and an acidic (Glu 84) residue in close spatial proximity that produce a characteristic electrostatic signature surrounding the region of Cys 61. Interestingly, in the one case (Fig. 5e) in which Glu 84 is not conserved, the model is predicted to present a unique acidic residue to the surface immediately adjacent to this site (to the right of the green circle in panel e), so that the overall electrostatic character surrounding Cys 61 is conserved. A PrISM multiple structure superposition of the YgdK structure and the four homology models reveal that 90% of their C^{α} backbone atoms are "within" an RMSD value of 2 Å, while only 9% are more diverse, sharing RMSD values in the range of 2–4 Å. This supports the view that our homology models are of high quality. Furthermore, the residues denoted in Figure 5 exhibit rather low fluctuations among the experimental structure and the four models (with average RMSD values of 0.5 Å for Cys 61, 1.0 Å for Glu 84, and 1.5 Å for Arg 129). Taken together, our findings strongly suggest that this surface patch of YgdK and its homologs is functionally important, likely for complex formation with CsdA and its homologs.

Since it has been shown that (1) SufE and SufS form a binary complex (Loiseau et al. 2003; Ollagnier-de-Choudens et al. 2003; Outten et al. 2003) and (2) YgdK and CsdA share, respectively, comparably high sequence homology to SufE and SufS, it is quite likely that YgdK forms a complex with CsdA analogous to the SufE/SufS complex. As a first step toward understanding the structural biology of the complex formation, we built a homology model of CsdA based on the X-ray crystal structure of SufS (PDB ID 1I29; Mihara et al. 2002), which exhibits 45% sequence identity with CsdA. Subsequently, the spatial clustering of residues conserved within the SufS/CsdA family of sequence homologs was analyzed using the program ConSurf (Glaser et al. 2003), and the conserved surface patches and surface electrostatic potentials of YgdK and CsdA were compared (Fig. 6). Based on their complementarity (green arrows), including the active-site cysteines (Cys 61 in YgdK and Cys 358 in CsdA), it is tempting to propose that complex formation would occur by rotating the YgdK structure, as shown in Figure 6, by ~180° about a vertical axis and laying the structure on top of the model. Then, the conserved cysteines would be in close apposition for sulfur transfer, and three conserved basic/acidic pairs (yellow arrows) would provide charge matching.

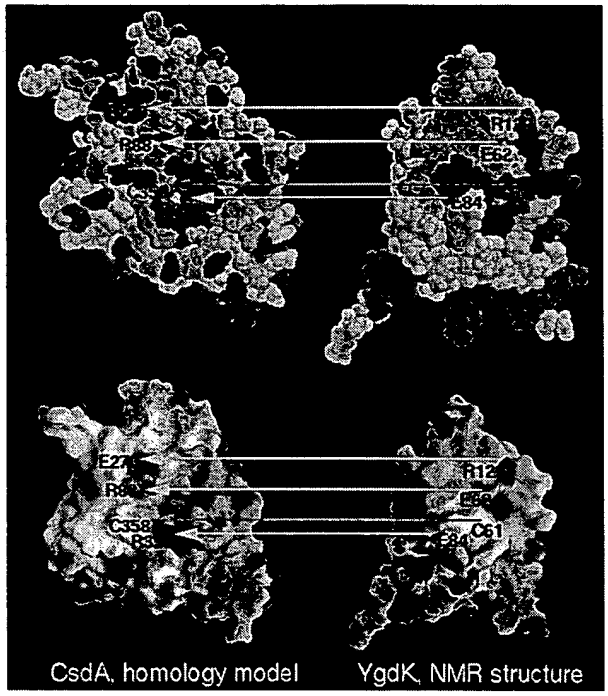


Figure 6. Proposed model of the interaction between CsdA and YgdK. A homology model of CsdA (shown on the *left*) was derived from the X-ray crystal structure of SufS (PDB ID: 1I29), with which it shares 45% sequence identity. The NMR structure of YgdK is shown on the right. The upper panel displays conserved surface residues inferred, respectively, from sequence conservation among the CsdA and YgdK families of sequence homologs using the program ConSurf (Glaser et al. 2003). Purple residues are most highly conserved, and cyan residues are least conserved. The lower panel displays surface electrostatic potential images calculated with the program GRASP (Nicholls et al. 1991), where red and blue surfaces denote negative and positive electrostatic potentials, respectively. The surfaces of CdsA and YgdK predicted to interact are facing the viewer. Based on the complementarity of conserved residues, in particular the active-site cysteines, as well as the electrostatic surface potentials, it is predicted that the YgdK/CsdA complex would be obtained by rotating the YgdK structure by ~180° about a vertical axis and laying it on top of the CsdA model.

In spite of lack of significant sequence identity, YgdK and IscU are structurally similar (Goldsmith-Fischman et al. 2004; Ramelot et al. 2004). IscU is homologous to the N-terminal domain of NifU, which contains a labile $[2Fe\text{-}2S]^{2+}$ binding site. It is thought that both NifU and IscU function as [Fe-S] cluster scaffold proteins, while there is no evidence to date that YgdK and SufS can function as [Fe-S] cluster scaffold proteins. In

contrast to IscU, which binds Zn under conditions of its structure determination (Goldsmith-Fischman et al. 2004; Ramelot et al. 2004), SufE and YgdK are not metallo-proteins. However, key features of the SufE–SufS or YdgK–CsdA interaction are similarly predicted for the IscU–IscS interaction (Goldsmith-Fischmann et al. 2004; Ramelot et al. 2004).

Materials and methods

Protein expression and purification

Uniformly (*U*) ^{13}C,^{15}N-labeled YgdK was expressed and purified following standard protocols (Acton et al. 2005). Briefly, the full-length YgdK gene from *E. coli* was cloned into a pET21d (Novagen) derivative, yielding the plasmid pER75-21. The resulting construct contains eight nonnative residues at the C terminus (LEHHHHHH) that facilitate protein purification. *E. coli* BL21 (DE3) pMGK cells, a rare codon-enhanced strain, were transformed with pER75-21, and cultured in MJ minimal medium (Jansson et al. 1996) containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as sole nitrogen and carbon sources. *U*-^{13}C,^{15}N YgdK was purified using a two-step protocol consisting of Ni-NTA affinity (QIAGEN) and gel filtration (HiLoad 26/60 Superdex 75; Amersham Biosciences) chromatography. The final yield of purified *U*-^{13}C,^{15}N YgdK (>97% homogeneous by SDS-PAGE; 17 kDa by MALDI-TOF mass spectrometry) was ~90 mg/L. An additional batch of *U*-^{15}N YgdK was prepared according to the same protocol, but using a minimal medium containing $(^{15}NH_4)_2SO_4$ and unlabeled glucose. The concentration of both stable isotope-labeled YgdK samples was 1 mM in 95% H_2O/5% D_2O solution containing 20 mM MES, 100 mM NaCl, 10 mM DTT, 5 mM $CaCl_2$, and 0.02% NaN_3 at pH 6.5.

NMR data acquisition and processing

NMR data were collected at 25°C on Varian INOVA 600 and 750 spectrometers. The spectra were processed and analyzed using the programs NMRPipe (Delaglio et al. 1995) and XEASY (Bartels et al. 1995), respectively. Resonance assignments were obtained as described (Szyperski et al. 2002) using a suite of reduced-dimensionality NMR experiments, including 3D HNNCAHA, HACA(CO)NHN, $H^{\alpha\beta}C^{\alpha\beta}$(CO)NHN, HCCH-COSY, and 2D HBCB(CGCD)HD and ^1H-TOCSY relayed HCH-COSY. These data were complemented by conventional (Cavanagh et al. 1996) 3D HNNCACB and HC(C)H TOCSY, and 3D HNNHA (Vuister and Bax 1993) for measurement of $^3J_{HN\alpha}$ couplings. Upper-limit distance constraints were extracted from 3D ^{15}N-resolved [^1H,^1H]-NOESY (Cavanagh et al. 1996) (τ_m = 70 msec) and ^{13}C-resolved [^1H,^1H]-NOESY (Cavanagh et al. 1996) (τ_m = 70 msec).

For combined analysis of conventional and RD NMR spectra using the program XEASY (Bartels et al. 1995), a suite of scripts was implemented to transfer chemical shifts into RD NMR peak lists, thereby recognizing the distinct peak pattern manifested in the various experiments (see Fig. 2 in Szyperski et al. 2002). Initially, peak lists for the RD NMR spectra with proposed resonance assignments were generated from (1) ^1HN and ^{15}N chemical shifts of spin systems identified in 2D [^{15}N,^1H] HSQC and (2) the ^1H and ^{13}C random coil values of chemical shifts measured in the projected dimension. The peak lists thus created were then manually adjusted. Once the assignment of the triple resonance spectra was (largely) completed, peak lists for the heteronuclear resolved NOESY spectra were created. These lists comprised intra, sequential, and medium range NOEs considering the protein's secondary structure as inferred from $^{13}C^{\alpha}$ chemical shifts, and were completed by manual peak picking.

NMR structure calculations

NOESY cross-peak volumes and $^3J_{HN\alpha}$ scalar coupling constants were converted into proton–proton upper distance limit and ϕ-angle constraints using the program DYANA (Güntert et al. 1997). Additional ϕ and ψ backbone dihedral angle constraints were derived from chemical shifts using the program TALOS (Cornilescu et al. 1999). The final round of DYANA structure calculations using torsion angle dynamics was started with 100 random conformers and 10,000 annealing steps (Güntert et al. 1997). The 20 structures with the lowest target functions were selected to represent the NMR solution structure. The calculation of RMSD values and solvent-exposed surface areas was performed using the program MOLMOL (Koradi et al. 1996).

Homology modeling and leverage analysis

PSI-BLAST (Altschul et al. 1997) searches against the nonredundant protein database were conducted to detect sequence homologs of YgdK and SufE. The BLOSUM62 substitution matrix (Henikoff and Henikoff 1992) was used with "gap existence" and "extension" penalties of 11 and 1, respectively. Using an inclusion *E*-value threshold of 0.001, the searches with both YgdK and SufE converged to the same set of sequences after three PSI-BLAST iterations. The sequences of YgdK and SufE and the homologs detected in the PSI-BLAST search were analyzed with the program PrISM (Yang and Honig 1999): (1) An all-on-all Needleman-Wunsch (global) sequence alignment provided the basis for pairwise sequence identities, the clustering of sequences into similar groups, and the construction of multiple sequence alignments, and (2) Smith-Waterman (local) sequence scans of the set of sequences homologs against the sequences for YgdK and SufE were used to determine which of the two structures provides the more suitable template for each homologous sequence. In cases in which a sequence chosen for modeling (the target sequence) had >30% sequence identity to the sequence of its structural template, the PrISM global alignment was used as the alignment for homology modeling. In cases in which the sequence identity between target and template was <30%, the alignment for modeling was extracted from a multiple alignment of all homologous sequences with the templates. The program NEST (Petrey et al. 2003) was used to construct all homology models. Each of the models was evaluated by the programs Verify 3D (Bowie et al. 1991; Luthy et al. 1992) and Prosa2 (Sippl 1993), which score structures according to how well each residue fits into its structural environment based on criteria derived from statistical analysis of high-resolution structures in the PDB.

Electronic supplemental material

The Supplemental Material contains a detailed description of results obtained from the search for proteins exhibiting structural similarity to YdgK.

Acknowledgments

This work was supported by the NIH (P50 GM62413-01) and the NSF (MCB 00075773 to T.S.). We thank G. Kornhaber for useful comments on the manuscript, and the UB Center of Computational Research (CCR) for support.

References

Acton, T.B., Gunsalus, K.C., Xiao, R., Ma, L.-C., Aramini, J.M., Baran, M.C., Chiang, Y.-W., Climent, T., Cooper, B., Denissova, N., et al. 2005. Robotic cloning and protein productions platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* **394:** 210–243.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aurora, R. and Rose, G.D. 1998. Helix capping. *Protein Sci.* **7:** 21–38.

Bartels, C., Xia, T.H., Billeter, M., Güntert, P., and Wüthrich, K. 1995. The program XEASY for computer-supported NMR spectral-analysis of biological macromolecules. *J. Biomol. NMR* **6:** 1–10.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Billeter, M. 1992. Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. *Q. Rev. Biophys.* **25:** 325–377.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253:** 164–169.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G., and Skelton, N.J. 1996. Heteronuclear NMR experiments. In *Protein NMR spectroscopy*, pp. 410–453. Wiley, New York.

Clore, G.M. and Gronenborn, A.M. 1998. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl Acad. Sci.* **95:** 5891–5898.

Cornilescu, G., Delaglio, F., and Bax, A. 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13:** 289–302.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. 1995. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6:** 277–293.

Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling. In *Computational biochemistry and biophysics* (eds. M. Watanabe et al.), pp. 275–312. Marcel Dekker, New York.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19:** 163–164.

Goldsmith-Fischman, S., Kuzin, A., Edstrom, W.C., Benach, J., Shastry, R., Xiao, R., Acton, T.B., Honig, B., Montelione, G.T., and Hunt, J.F. 2004. The SufE sulfur-acceptor protein contains a conserved core structure that mediates interdomain interactions in a variety of redox protein complexes. *J. Mol. Biol.* **344:** 549–565.

Gronenborn, A.M. and Clore, G.M. 1994. Identification of N-terminal helix capping boxes by means of ^{13}C chemical-shifts. *J. Biomol. NMR* **4:** 455–458.

Güntert, P., Mumenthaler, C., and Wüthrich, K. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273:** 283–298.

Harper, E.T. and Rose, G.D. 1993. Helix stop signals in proteins and peptides: The capping box. *Biochemistry* **32:** 7605–7609.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Holm, L. and Sander, C. 1995. Dali: A network tool for protein structure comparison. *Trends Biochem. Sci.* **20:** 478–480.

Jansson, M., Li, Y.-C., Jendeberg, L., Anderson, S., Montelione, G.T., and Nilsson, B. 1996. High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in *Escherichia coli*. *J. Biomol. NMR* **7:** 131–141.

Koradi, R., Billeter, M., and Wüthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14:** 51–55.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26:** 283–291.

Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of proteins structures solved by NMR. *J. Biomol. NMR* **8:** 477–486.

Liu, J. and Rost, B. 2002. Target space for structural genomics revisited. *Bioinformatics* **18:** 922–933.

Liu, J., Hegyi, H., Acton, T.B., and Montelione, G.T. 2004. Automatic target selection for structural genomics on eukaryotes. *Proteins* **56:** 188–200.

Loiseau, L., Ollangnier-de-Choudens, S., Nachin, L., Fontecave, M., and Barras, F. 2003. Biogenesis of Fe-S cluster by the bacterial Suf system. *J. Biol. Chem.* **278:** 38352–38359.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* **356:** 83–85.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 291–325.

Mihara, H. and Esaki, N. 2002. Bacterial cysteine desulfurases: Their function and mechanisms. *Appl. Microbiol. Biotechnol.* **60:** 12–23.

Mihara, H., Fujii, T., Kato, S., Kurihara, T., Hata, Y., and Esaki, N. 2002. Structure of external aldimine of *Escherichia coli* CsdB, an IscS/NifS homolog: Implications for its specificity toward selenocysteine. *J. Biochem.* **131:** 679–685.

Nicholls, A., Sharp, K., and Honig, B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11:** 281–296.

Ollagnier-de-Choudens, S., Lascoux, D., Loiseau, L., Barras, F., Forest, E., and Fontecave, M. 2003. Mechanistic studies of the SufS–SufE cysteine desulfurase: Evidence for sulfur transfer from SufS to SufE. *FEBS Lett.* **555:** 263–267.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Outten, F.W., Wood, M.J., Munoz, F.M., and Storz, G. 2003. The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in *Escherichia coli*. *J. Biol. Chem.* **278:** 45713–45719.

Palmer, A.G. 2001. NMR probes of molecular dynamics: Overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **30:** 129–155.

Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., and Orengo, C.A. 2000. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28:** 277–282.

Petrey, D., Xiang, X., Tang, C., Xie, L., and Gimpelev, M. 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* **53:** 430–435.

Ramelot, T.A., Cort, J.R., Goldsmith-Fischman, S., Kornhaber, G.J., Xiao, R., Shastry, R., Acton, T.B., Montelione, G.T., and Kennedy, M.A. 2004. Solution NMR structure of the iron-sulfur cluster assembly protein U (IscU) with zinc bound at the active site. *J. Mol. Biol.* **344:** 567–583.

Schwede, T., Diemand, A., Guex, N., and Peitsch, M.C. 2000. Protein structure computing in the genomic era. *Res. Microbiol.* **151:** 107–112.

Shindyalov, I.N. and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11:** 739–747.

Sippl, M.J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17:** 355–362.

Smith, A.D., Agar, J.N., Johnson, K.A., Frazzon, J., Amster, I.J., Dean, D.R., and Johnson, M.K. 2001. Sulfur transfer from IscS to IscU: The first step in iron-sulfur cluster biosynthesis. *J. Am. Chem. Soc.* **123:** 11103–11104.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N.B., and Montelione, G.T. 2002. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. *Proc. Natl Acad. Sci.* **99:** 8009–8014.

Tramontano, A. and Morea, V. 2003. Assessment of homology-based predictions in CASP5. *Proteins* **53:** 352–386.

Urbina, H.D., Silberg, J.J., Hoff, K.G., and Vickery, L.E. 2001. Transfer of sulfur from IscS to IscU during Fe/S cluster assembly. *J. Biol. Chem.* **276:** 44521–44526.

Vuister, G.W. and Bax, A. 1993. Quantitative J correlation: A new approach for measuring homonuclear three-bond $J(H^NH^\alpha)$ coupling constants in ^{15}N-enriched proteins. *J. Am. Chem. Soc.* **115:** 7772–7777.

Wunderlich, Z., Acton, T.B., Liu, J., Kornhaber, G., Everett, J., Carter, P., Lan, N., Echols, N., Gerstein, M., Rost, B., et al. 2004. The protein target list of the Northeast Structural Genomics Consortium. *Proteins* **56:** 181–187.

Wüthrich, K. 1986. NOE-observable ^{1}H-^{1}H distances in proteins. In *NMR of proteins and nucleic acids*, pp. 117–129. Wiley, New York.

Yang, A.S. and Honig, B. 1999. Sequence to structure alignment in comparative modeling using PrISM. *Proteins* **3(Suppl.):** 66–72.